TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**LÊ VĂN CƯỜNG - 52100171**

**TỐI ƯU HÓA CHIẾN LƯỢC KINH DOANH QUA PHÂN TÍCH DỮ LIỆU VÀ DỰ ĐOÁN HÀNH VI KHÁCH HÀNG**

**BÁO CÁO CUỐI KỲ**

**HỆ THỐNG THƯƠNG MẠI**

**THÔNG MINH**

Người hướng dẫn:

**ThS. Dương Hữu Phúc**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

# MỤC LỤC

[MỤC LỤC ii](#_Toc3951)

[DANH MỤC HÌNH VẼ vii](#_Toc17465)

[DANH MỤC BẢNG BIỂU ix](#_Toc26095)

[CHƯƠNG 1. PHÁT BIỂU ĐỀ TÀI 1](#_Toc32320)

[1.1 Bối cảnh 1](#_Toc17373)

[1.1.1 Tổng quan về ngành bán lẻ 1](#_Toc29316)

[1.1.2 Thách thức trong việc “Nâng cao khả năng quay lại của khách hàng” 1](#_Toc9865)

[1.2 Tầm quan trọng 2](#_Toc20287)

[1.2.1 Duy trì và tăng trưởng doanh thu 2](#_Toc14529)

[1.2.2 Tăng cường lòng trung thành và giá trị thương hiệu 2](#_Toc31612)

[1.2.3 Tối ưu hóa chiến lược kinh doanh 3](#_Toc30455)

[1.2.4 Cạnh tranh trên thị trường 3](#_Toc13620)

[1.3 Vấn đề 3](#_Toc13664)

[1.3.1 Tỷ lệ khách hàng quay lại thấp 3](#_Toc23902)

[1.3.2 Thiếu hiệu quả trong phân tích dữ liệu khách hàng 3](#_Toc8269)

[1.3.3 Chương trình khuyến mãi chưa đủ hấp dẫn 4](#_Toc18805)

[1.3.4 Khó khăn trong việc đưa ra quyết định dựa trên dữ liệu 4](#_Toc20494)

[1.4 Mục tiêu 4](#_Toc28227)

[1.4.1 Tối ưu hóa chương trình khuyến mãi 4](#_Toc5400)

[1.4.2 Nâng cao hiệu quả phân tích dữ liệu khách hàng 4](#_Toc5616)

[1.4.3 Hỗ trợ ra quyết định 4](#_Toc1615)

[1.5 Phạm vi 5](#_Toc9132)

[1.5.1 Phạm vi về dữ liệu 5](#_Toc27292)

[1.5.2 Phạm vi về phân tích 5](#_Toc14691)

[1.5.3 Phạm vi đối tượng 5](#_Toc8251)

[1.6 Kết quả 5](#_Toc30814)

[CHƯƠNG 2. ĐẶC TẢ DỮ LIỆU 7](#_Toc30560)

[2.1 Giới thiệu 7](#_Toc7115)

[2.2 Đặc điểm dữ liệu 9](#_Toc25383)

[2.2.1  Thông tin tổng quan 9](#_Toc25712)

[2.2.2 Các thông số thống kê cơ bản 10](#_Toc15085)

[2.2.3 Đánh giá tính chất dữ liệu 11](#_Toc18206)

[2.3 Các đặc tính của dữ liệu 11](#_Toc22401)

[2.3.1 Đặc tính về loại dữ liệu 11](#_Toc7074)

[2.3.2 Đặc tính về cấu trúc dữ liệu 12](#_Toc29394)

[2.3.3 Đặc tính về giá trị dữ liệu 12](#_Toc14396)

[2.3.4 Đặc tính về phân phối dữ liệu 13](#_Toc9724)

[2.3.5 Đặc tính về mối quan hệ dữ liệu 13](#_Toc1520)

[2.3.6 Đặc tính về tính nhạy cảm 13](#_Toc235)

[2.3.7 Đặc tính về tính động 13](#_Toc15395)

[2.3.8 Đặc tính về khả năng sử dụng 14](#_Toc20461)

[CHƯƠNG 3. TIỀN XỬ LÝ DỮ LIỆU 15](#_Toc9831)

[3.1 Xử lý các ký hiệu đặc biệt và dữ liệu dạng số 15](#_Toc1763)

[3.2 Kiểm tra xử lý dữ liệu bị thiếu (Missing values) 15](#_Toc11619)

[3.3 Xử lý giá trị trùng lặp (Duplicates) 15](#_Toc18165)

[CHƯƠNG 4. TRỰC QUAN HÓA DỮ LIỆU 16](#_Toc25656)

[4.1 Doanh thu theo thời gian 16](#_Toc23288)

[4.1.1 Tổng quan xu hướng 16](#_Toc9683)

[4.1.2 Các điểm nổi bật 16](#_Toc10209)

[4.1.3 Đặc điểm chu kỳ 16](#_Toc27147)

[4.1.4 Kết luận 17](#_Toc14846)

[4.2 Phân phối độ tuổi khách hàng 17](#_Toc30849)

[4.2.1 Tổng quan 17](#_Toc7374)

[4.2.2 Chi tiết phân phối 18](#_Toc3989)

[4.2.3 Đặc điểm nổi bật 18](#_Toc23207)

[4.2.4 Kết luận 18](#_Toc2031)

[4.3 Doanh thu theo mục sản phẩm 19](#_Toc10897)

[4.3.1 Tổng quan 19](#_Toc365)

[4.3.2 Chi tiết từng danh mục 19](#_Toc17921)

[4.3.3 Những điểm nổi bật 20](#_Toc19580)

[4.3.4 Kết luận 20](#_Toc27990)

[4.4 Tỷ lệ đơn hàng theo phương thức thanh toán 21](#_Toc6939)

[4.4.1 Phương thức thanh toán phổ biến nhất 21](#_Toc6464)

[4.4.2 Phương thức thanh toán phổ biến thứ hai 22](#_Toc7968)

[4.4.3 Các phương thức khác 22](#_Toc2182)

[4.4.4 Phương thức ít phổ biến nhất 22](#_Toc10098)

[4.4.5 Nhận định tổng quan 22](#_Toc5310)

[4.4.6 Đề xuất 22](#_Toc7260)

[4.5 Số lượng khách hàng theo khu vực 23](#_Toc19951)

[4.6 Tổng doanh thu theo khu vực 24](#_Toc31521)

[4.6.1 Khu vực có doanh thu cao nhất 24](#_Toc7232)

[4.6.2 Khu vực có doanh thu thấp nhất 24](#_Toc13169)

[4.6.3 Tương quan giữa các khu vực 24](#_Toc24607)

[4.6.4 Nhận định tổng quan 25](#_Toc7556)

[4.7 Tỉ lệ giảm giá trung bình theo danh mục 25](#_Toc3195)

[4.7.1 Danh mục có tỷ lệ giảm giá cao nhất 25](#_Toc13777)

[4.7.2 Danh mục có tỷ lệ giảm giá thấp 25](#_Toc21520)

[4.7.3 Sự chênh lệch trong tỷ lệ giảm giá 26](#_Toc29632)

[4.7.4 Ý nghĩa trong kinh doanh 26](#_Toc2519)

[CHƯƠNG 5. CÔNG NGHỆ VÀ PHƯƠNG PHÁP 27](#_Toc31443)

[5.1 Công nghệ 27](#_Toc5643)

[5.1.1 Python 27](#_Toc8718)

[5.1.2 Chartjs 27](#_Toc11378)

[5.1.3 Flask 27](#_Toc27562)

[5.1.4 TensorFlow 27](#_Toc5574)

[5.1.5 Mongodb 28](#_Toc5975)

[5.1.6 Tableau 28](#_Toc11624)

[5.2 Phương pháp 28](#_Toc2968)

[5.2.1 Phân cụm khách hàng 28](#_Toc32208)

[5.2.2 Dự đoán khách hàng quay lại 29](#_Toc30466)

[5.2.3 Dự đoán số lượng khách hàng theo chuỗi thời gian 29](#_Toc31058)

[5.3 Kết quả 30](#_Toc20575)

[5.4 Ứng dụng 30](#_Toc31278)

[CHƯƠNG 6. GIẢI QUYẾT VẤN ĐỀ 31](#_Toc21770)

[6.1 Phân Cụm Khách Hàng theo RFM bằng KMeans 31](#_Toc19468)

[6.1.1 Cách Thực Hiện Phân Cụm Khách Hàng theo RFM bằng KMeans 33](#_Toc24313)

[6.1.2 Lợi Ích Của Phân Cụm Khách Hàng theo RFM 36](#_Toc21063)

[6.2 Dự Đoán Sự Trở Lại Mua Hàng của Khách Hàng (Random Forest) 37](#_Toc30803)

[6.2.1 Giới Thiệu về Random Forest 37](#_Toc14780)

[6.2.2 Nguyên Lý Hoạt Động của Random Forest 38](#_Toc136)

[6.2.3 Các Yếu Tố Ảnh Hưởng Đến Dự Đoán Sự Quay Lại Mua Hàng 39](#_Toc23037)

[6.2.4 Các Bước Tiền Xử Lý Dữ Liệu 40](#_Toc5950)

[6.2.5 Xây Dựng Mô Hình Dự Đoán 40](#_Toc18703)

[6.2.6 Trả Về Phần Trăm Quay Lại của Mỗi Khách Hàng 42](#_Toc25488)

[6.2.7 Lợi Ích của Việc Sử Dụng Random Forest 44](#_Toc32285)

[6.3 Dự Đoán Số Lượng Khách Hàng Theo Ngày (RNN) 45](#_Toc17019)

[6.3.1 Giới Thiệu về RNN 45](#_Toc710)

[6.3.2 Cấu Trúc và Cách Hoạt Động của RNN 46](#_Toc32100)

[6.3.3 Cách Thực Hiện Dự Đoán Số Lượng Khách Hàng Theo Ngày bằng RNN 47](#_Toc28665)

[6.3.4 Lợi Ích Của Dự Đoán Số Lượng Khách Hàng Theo Ngày Bằng RNN 50](#_Toc29214)

[CHƯƠNG 7. ỨNG DỤNG WEB VÀ KẾT QUẢ ĐẦU RA 52](#_Toc6770)

[7.1 Dashboard 52](#_Toc8158)

[7.1.1 Month-Wise Revenue (Doanh thu theo tháng) 52](#_Toc13395)

[7.1.2 Age-Wise Sales Analysis (Phân tích doanh thu theo độ tuổi) 53](#_Toc26869)

[7.1.3 Revenue per State & Region-Wise Revenue Share (Doanh thu theo bang và khu vực) 54](#_Toc5243)

[7.1.4 Gender-Wise Sales Analysis (Phân tích doanh thu theo giới tính) 55](#_Toc5971)

[7.1.5 Quantity-Discount Correlation (Tương quan chiết khấu và lượng đặt hàng) 56](#_Toc6525)

[7.1.6 Kết luận 57](#_Toc25113)

[7.2 Predicting Returning Customers 57](#_Toc3431)

[7.2.1 Phân tích tỷ lệ quay lại theo các đặc điểm khách hàng 58](#_Toc25798)

[7.2.2 Top các yếu tố ảnh hưởng đến hành vi rời đi 59](#_Toc22719)

[7.2.3 Phân nhóm khách hàng có khả năng rời đi cao và thấp 60](#_Toc15556)

[7.2.4 Ý nghĩa của hệ thống đối với doanh nghiệp 62](#_Toc1625)

[7.3 Dự đoán số lượng khách hàng 62](#_Toc26716)

[7.3.1 Các chức năng chính của trang web 62](#_Toc26361)

[7.3.2 Kết quả đầu ra của mô hình RNN 63](#_Toc6754)

[7.3.3 Ý nghĩa đối với doanh nghiệp 65](#_Toc23256)

[7.4 Phân khúc khách hàng 66](#_Toc4112)

[7.4.1 Các chức năng chính của trang web 66](#_Toc14473)

[7.4.2 Kết quả đầu ra của mô hình K-Means 70](#_Toc8939)

[7.4.3 Ý nghĩa của phân khúc khách hàng đối với doanh nghiệp 72](#_Toc17362)

[7.4.4 Kết luận 73](#_Toc24232)

[TÀI LIỆU THAM KHẢO 74](#_Toc31855)

# DANH MỤC HÌNH VẼ

[Hình 4.1 : Doanh thu theo thời gian 17](#_Toc20465)

[Hình 4.2 : Phân phối độ tuổi khách hàng 19](#_Toc4612)

[Hình 4.3 : Tương quan Số lượng - Giảm giá 21](#_Toc30600)

[Hình 4.4 : Tỷ lệ đơn hàng theo phương thức thanh toán 23](#_Toc30069)

[Hình 4.5 : Số lượng khách hàng theo khu vực 24](#_Toc2027)

[Hình 4.6 : Tổng doanh thu theo khu vực 25](#_Toc1660)

[Hình 4.7 : Tỉ lệ giảm giá trung bình theo danh mục 26](#_Toc29695)

[Hình 7.1 : Doanh thu theo tháng 53](#_Toc28094)

[Hình 7.2 : Phân tích bán hàng theo độ tuổi 54](#_Toc15668)

[Hình 7.3 : Doanh thu trên mỗi tiểu bang 55](#_Toc26820)

[Hình 7.4 : Phân tích bán hàng theo giới tính 56](#_Toc17159)

[Hình 7.5 : Tương quan Số lượng - Giảm giá 57](#_Toc13731)

[Hình 7.6 : Predicting Returning Customers 58](#_Toc11212)

[Hình 7.7 : Phân tích tỷ lệ quay lại theo các đặc điểm khách hàng 59](#_Toc25052)

[Hình 7.8 : Top các yếu tố ảnh hưởng đến hành vi rời đi 60](#_Toc19114)

[Hình 7.9 : Phân nhóm khách hàng có khả năng rời đi trên 80% 61](#_Toc129)

[Hình 7.10 : Phân nhóm khách hàng có khả năng rời đi dưới 20% 61](#_Toc9959)

[Hình 7.11 : Dự đoán số lượng khách hàng của ngày tiếp theo 63](#_Toc2969)

[Hình 7.12 : Phân tích khách hàng theo tần suất và thời gian 67](#_Toc2149)

[Hình 7.13 : Tỷ lệ các nhóm khách hàng 67](#_Toc29894)

[Hình 7.14 : So sánh các giá trị trung bình các nhóm khách hàng 68](#_Toc11629)

[Hình 7.15 : So sánh các giá trị trung bình các nhóm khách hàng 68](#_Toc10533)

[Hình 7.16 : So sánh các giá trị trung bình các nhóm khách hàng 69](#_Toc3500)

[Hình 7.17 : Phân loại và phân khúc khách hàng 70](#_Toc3190)

# DANH MỤC BẢNG BIỂU

[Bảng 2.1 : Danh sách các thuộc tính đơn hàng 7](#_Toc20091)

# PHÁT BIỂU ĐỀ TÀI

## Bối cảnh

### Tổng quan về ngành bán lẻ

Ngành bán lẻ là một trong những lĩnh vực kinh tế quan trọng, đóng vai trò trực tiếp trong việc cung cấp sản phẩm và dịch vụ tới tay người tiêu dùng.

Ngành này bao gồm một loạt các hoạt động từ việc lưu trữ, phân phối đến việc bán lẻ các sản phẩm đa dạng như: thực phẩm, đồ gia dụng, quần áo đến công nghệ và dịch vụ.

Với sự phát triển mạnh mẽ của thương mại điện tử và các nền tảng trực tuyến, ngành bán lẻ đã trải qua những thay đổi lớn, từ cách thức kinh doanh truyền thống sang hình thức mua sắm trực tuyến, việc thay đổi mạnh mẽ cách thức hoạt động của ngành bán lẻ làm tăng thêm sự cạnh tranh trong thị trường.

Điều này đã dẫn đến nhu cầu ngày càng tăng về việc sử dụng các công cụ phân tích dữ liệu tiên tiến để hỗ trợ ra quyết định và cải thiện hiệu suất kinh doanh.

### Thách thức trong việc “Nâng cao khả năng quay lại của khách hàng”

Trong bối cảnh cạnh tranh ngày càng gay gắt, việc giữ chân khách hàng và khuyến khích họ quay lại mua sắm là một trong những thách thức lớn nhất đối với doanh nghiệp.

***Cạnh tranh cao từ thị trường:*** Sự bùng nổ của thương mại điện tử và sự xuất hiện của nhiều nhà bán lẻ mới tạo ra một môi trường cạnh tranh khốc liệt. Khách hàng có nhiều lựa chọn hơn và việc thu hút cũng như giữ chân họ trở nên khó khăn hơn.

***Thay đổi hành vi mua sắm:*** Hành vi mua sắm của người tiêu dùng đang thay đổi nhanh chóng với sự phát triển của công nghệ và sự phổ biến của các thiết bị di động. Khách hàng ngày càng có xu hướng mua sắm trực tuyến, đòi hỏi doanh nghiệp phải điều chỉnh chiến lược kinh doanh để đáp ứng nhu cầu này.

***Kỳ vọng cao từ khách hàng:*** Khách hàng hiện nay có kỳ vọng cao về chất lượng sản phẩm, dịch vụ khách hàng, và trải nghiệm mua sắm. Do đó, các doanh nghiệp cần cung cấp dịch vụ tốt hơn và cá nhân hóa trải nghiệm mua sắm để giữ chân khách hàng.

***Khó khăn trong việc phân tích dữ liệu:*** Với khối lượng dữ liệu khổng lồ được tạo ra hàng ngày, doanh nghiệp gặp khó khăn trong việc thu thập, phân tích và sử dụng dữ liệu để đưa ra quyết định kinh doanh hiệu quả.

***Chiến lược khách hàng thân thiết:*** Phát triển và duy trì chương trình khách hàng thân thiết hiệu quả là một thách thức lớn. Các chương trình này cần phải mang lại giá trị thực sự cho khách hàng, đồng thời giúp doanh nghiệp thu thập thông tin hữu ích về hành vi mua sắm của họ.

## Tầm quan trọng

Trong ngành bán lẻ, việc “nâng cao khả năng quay lại mua sản phẩm của khách hàng” không chỉ là một mục tiêu chiến lược mà còn là yếu tố sống còn để duy trì và phát triển doanh nghiệp.

### Duy trì và tăng trưởng doanh thu

Khách hàng quay lại mua sắm đóng góp trực tiếp vào doanh thu của doanh nghiệp. Các nghiên cứu cho thấy rằng việc giữ chân khách hàng hiện tại có chi phí thấp hơn rất nhiều so với việc thu hút khách hàng mới. Thậm chí, khách hàng quen thuộc thường chi tiêu nhiều hơn và có xu hướng mua sắm thường xuyên hơn, từ đó tạo ra nguồn doanh thu ổn định và gia tăng cho doanh nghiệp.

### Tăng cường lòng trung thành và giá trị thương hiệu

Lòng trung thành của khách hàng không chỉ đảm bảo doanh thu mà còn củng cố giá trị thương hiệu. Khi khách hàng cảm thấy hài lòng với trải nghiệm mua sắm, họ có xu hướng chia sẻ trải nghiệm đó với người khác, tạo ra hiệu ứng lan truyền tích cực. Điều này giúp tăng cường hình ảnh thương hiệu và thu hút thêm khách hàng mới thông qua sự giới thiệu từ những người tiêu dùng hiện tại.

### Tối ưu hóa chiến lược kinh doanh

Việc phân tích dữ liệu về hành vi mua sắm của khách hàng quay lại cho phép các doanh nghiệp bán lẻ tối ưu hóa chiến lược kinh doanh của mình.

Hiểu rõ về nhu cầu, sở thích và xu hướng của khách hàng giúp doanh nghiệp có thể điều chỉnh các chiến dịch tiếp thị, quản lý hàng tồn kho, và phát triển các sản phẩm hoặc dịch vụ mới phù hợp với thị hiếu của thị trường.

### Cạnh tranh trên thị trường

Trong môi trường cạnh tranh ngày càng gay gắt, việc giữ chân khách hàng và làm cho họ quay lại mua sắm trở thành một lợi thế cạnh tranh quan trọng. Doanh nghiệp nào có thể tối ưu hóa trải nghiệm khách hàng và tạo ra sự khác biệt trong dịch vụ sẽ dễ dàng vượt qua các đối thủ và chiếm lĩnh thị phần.

## Vấn đề

### Tỷ lệ khách hàng quay lại thấp

Hiện tại, chuỗi bán lẻ TCT đang đối mặt với thách thức lớn là tỷ lệ khách hàng quay lại mua sắm không cao. Điều này không chỉ ảnh hưởng đến doanh thu mà còn giảm khả năng xây dựng một cơ sở khách hàng trung thành.

Việc không giữ được khách hàng sau lần mua sắm đầu tiên là dấu hiệu của trải nghiệm khách hàng không tối ưu, hoặc chương trình khách hàng thân thiết chưa hiệu quả.

### Thiếu hiệu quả trong phân tích dữ liệu khách hàng

Chuỗi bán lẻ TCT hiện đang gặp khó khăn trong việc thu thập, phân tích và sử dụng dữ liệu khách hàng để đưa ra các quyết định chiến lược kinh doanh. Dữ liệu từ các hoạt động mua sắm chưa được khai thác tối đa để hiểu rõ hơn về hành vi, nhu cầu và xu hướng của khách hàng.

Việc thiếu khả năng phân tích dữ liệu làm hạn chế khả năng đưa ra quyết định, đánh giá liên quan đến chiến lược kinh doanh.

### Chương trình khuyến mãi chưa đủ hấp dẫn

Chưa đề xuất được chương trình khuyến mãi phù hợp, hấp dẫn đối với từng đối tượng khách hàng hay từng loại sản phẩm, dẫn đến việc khách hàng không cảm thấy hứng thú để quay lại mua sắm.

Chương trình này cần được tối ưu hóa dựa trên phân tích dữ liệu giao dịch như thông tin khách hàng, thông tin giỏ hàng để đảm bảo nó thực sự mang lại giá trị cho khách hàng, từ đó nâng cao tỷ lệ quay lại mua sắm.

### Khó khăn trong việc đưa ra quyết định dựa trên dữ liệu

Việc thiếu một hệ thống Business Intelligence mạnh mẽ khiến cho việc đưa ra các quyết định chiến lược dựa trên dữ liệu gặp nhiều khó khăn. Điều này dẫn đến sự thiếu hiệu quả trong việc điều chỉnh các chiến dịch tiếp thị và cải thiện dịch vụ khách hàng.

## Mục tiêu

### Tối ưu hóa chương trình khuyến mãi

Mục tiêu chính của dự án là tăng tỷ lệ khách hàng quay lại mua sắm tại chuỗi bán lẻ TCT. Điều này sẽ được thực hiện bằng cách đưa ra các chương trình khuyến mãi phù hợp, hấp dẫn từ đó gia tăng khả năng quay lại của khách hàng.

Dự án sẽ phân tích và đánh giá về xu hướng mua sắm, phân khúc khách hàng để có thể đề xuất chương trình khuyến mãi phù hợp.

### Nâng cao hiệu quả phân tích dữ liệu khách hàng

Xây dựng và triển khai các công cụ Business Intelligence để phân tích dữ liệu một cách hiệu quả hơn. Dữ liệu này sẽ được sử dụng để hiểu rõ hơn về hành vi mua sắm của khách hàng, từ đó hỗ trợ các quyết định chiến lược kinh doanh.

### Hỗ trợ ra quyết định

Cung cấp các mô hình học máy, các công cụ BI, ứng dụng để tương tác với dữ liệu nhằm hỗ trợ để hỗ việc ra các quyết định kinh doanh.

## Phạm vi

### Phạm vi về dữ liệu

***Nguồn dữ liệu:*** Dự án sẽ tập trung vào việc thu thập và phân tích dữ liệu từ hệ thống bán hàng của chuỗi bán lẻ TCT.

***Loại dữ liệu:*** Dữ liệu bao gồm một số thuộc tính quan trọng như: thông tin giao dịch, thông tin cá nhân khách hàng.

***Thời gian dữ liệu:*** Dữ liệu được sử dụng sẽ bao gồm khoảng thời gian từ 1 đến 2 năm trước khi dự án bắt đầu để có đủ thông tin phục vụ cho phân tích và dự báo.

### Phạm vi về phân tích

***Phân tích hành vi khách hàng:*** Dự án sẽ phân tích các yếu tố ảnh hưởng đến hành vi quay lại của khách hàng, bao gồm cả yếu tố nhân khẩu học, tần suất mua sắm, giá trị đơn hàng trung bình.

### Phạm vi đối tượng

***Khách hàng:*** Dự án tập trung vào việc phân tích dữ liệu giao dịch của các khách hàng.

**Doanh nghiệp:** Cung cấp cho nhân viên trong doanh nghiệp bao gồm các chuyên gia về BI, phân tích dữ liệu, tiếp thị và quản lý khách hàng.

## Kết quả

Dự án sẽ cung cấp cho doanh nghiệp một bộ công cụ phân tích dữ liệu và hỗ trợ ra quyết định, giúp tối ưu hóa chiến lược kinh doanh nhằm tăng cường tỷ lệ khách hàng quay lại mua sắm. Các kết quả cụ thể dự kiến đạt được bao gồm:

**Công cụ phân tích và phân khúc khách hàng:** Xây dựng các mô hình phân tích giúp doanh nghiệp hiểu rõ hơn về hành vi mua sắm của khách hàng, từ đó phân chia khách hàng thành các nhóm cụ thể dựa trên các tiêu chí như tần suất mua sắm, giá trị đơn hàng trung bình, và các yếu tố nhân khẩu học. Điều này sẽ giúp doanh nghiệp điều chỉnh chiến lược tiếp thị và khuyến mãi sao cho phù hợp với từng nhóm khách hàng.

**Dự đoán tỷ lệ khách hàng quay lại:** Phân tích và dự báo tỷ lệ khách hàng quay lại mua sắm trong tương lai, giúp doanh nghiệp đánh giá hiệu quả của các chiến lược hiện tại và đưa ra các biện pháp cải thiện.

**Dự đoán số lượng khách hàng hàng ngày tại cửa hàng:** Xây dựng mô hình dự báo số lượng khách hàng đến mua sắm tại từng cửa hàng trong một khoảng thời gian nhất định, giúp doanh nghiệp tối ưu hóa kế hoạch nhân sự và quản lý kho hàng hiệu quả hơn.

**Cải thiện khả năng ra quyết định:** Cung cấp các công cụ Business Intelligence (BI) mạnh mẽ giúp đội ngũ quản lý, chuyên gia tiếp thị và phân tích dữ liệu đưa ra các quyết định chiến lược kịp thời và chính xác, từ đó nâng cao hiệu quả kinh doanh và tăng trưởng bền vững

Thông qua việc triển khai các công cụ phân tích và dự báo này, doanh nghiệp sẽ có cái nhìn rõ ràng hơn về hành vi khách hàng, từ đó có thể tối ưu hóa các chiến lược kinh doanh, tăng cường khả năng giữ chân khách hàng và nâng cao doanh thu.

# ĐẶC TẢ DỮ LIỆU

## Giới thiệu

Là bộ dữ liệu Sales Data FY 2020-2021 được lấy từ Kaggle chứa thông tin về các đơn đặt hàng của khách hàng trên trang web thương mại điện tử. Nó bao gồm các chi tiết về các đơn đặt hàng như ID đơn hàng, ngày đặt hàng, trạng thái đơn hàng, ID mặt hàng, số lượng đặt hàng, giá cả, giá trị, số tiền chiết khấu, tổng số, ID khách hàng, năm, số tham chiếu, tuổi, mã zip và tỷ lệ chiết khấu.

Tập dữ liệu có tổng cộng 286.392 mục với 35 cột. ID mặt hàng, số lượng đặt hàng, giá, giá trị, số tiền chiết khấu, tổng số, ID khách hàng, năm, số tham chiếu, tuổi, mã zip và phần trăm chiết khấu là dữ liệu số, trong khi ID đơn hàng, ngày đặt hàng và trạng thái đơn hàng là dữ liệu phân loại.

Bảng 2.1: Danh sách các thuộc tính đơn hàng

|  |  |
| --- | --- |
| **Tên Trường** | **Ý Nghĩa** |
| order\_id | Mã đơn hàng duy nhất. |
| order\_date | Ngày đặt hàng. |
| item\_id | Mã sản phẩm. |
| sku | Mã SKU (Stock Keeping Unit) của sản phẩm. |
| qty\_ordered | Số lượng sản phẩm đặt mua. |
| price | Giá của mỗi sản phẩm. |
| value | Tổng giá trị sản phẩm trước giảm giá. |
| discount\_amount | Số tiền giảm giá. |
| total | Tổng số tiền sau giảm giá. |
| category | Danh mục sản phẩm. |
| payment\_method | Phương thức thanh toán. |
| bi\_st | Trạng thái kinh doanh. |
| cust\_id | ID khách hàng. |
| year | Năm đặt hàng. |
| month | Tháng đặt hàng (kèm năm). |
| ref\_num | Số tham chiếu liên quan đến đơn hàng. |
| Name Prefix | Tiền tố tên của khách hàng. |
| First Name | Tên khách hàng. |
| Middle Initial | Chữ cái đầu của tên đệm. |
| Last Name | Họ của khách hàng. |
| Gender | Giới tính của khách hàng. |
| age | Tuổi của khách hàng. |
| full\_name | Họ và tên đầy đủ của khách hàng (tự động ghép từ họ, tên). |
| E Mail | Địa chỉ email của khách hàng. |
| Customer Since | Ngày khách hàng đăng ký lần đầu. |
| SSN | Mã số an sinh xã hội của khách hàng (nhạy cảm). |
| Phone No. | Số điện thoại của khách hàng. |
| Place Name | Tên địa điểm. |
| County | Quận của địa điểm. |
| City | Thành phố của địa điểm. |
| State | Bang của địa điểm. |
| Zip | Mã bưu chính. |
| Region | Vùng miền của địa điểm. |
| User Name | Tên đăng nhập của khách hàng. |
| Discount\_Percent | Phần trăm giảm giá áp dụng. |

## Đặc điểm dữ liệu

### Thông tin tổng quan

**Thông tin đơn hàng**:

* order\_id: Mã đơn hàng (định danh duy nhất).
* order\_date: Ngày đặt hàng.
* item\_id, sku: Mã sản phẩm.
* qty\_ordered: Số lượng sản phẩm đặt.
* price: Giá sản phẩm.
* value: Giá trị của sản phẩm (số lượng × giá).
* discount\_amount: Số tiền giảm giá.
* total: Tổng tiền của mục chi tiết (đã áp dụng giảm giá).

**Thông tin khách hàng**:

* cust\_id: Mã khách hàng.
* Name Prefix, First Name, Middle Initial, Last Name: Họ và tên khách hàng.
* Gender: Giới tính.
* age: Tuổi.
* full\_name: Tên đầy đủ của khách hàng (kết hợp từ các thành phần tên).
* E Mail: Email của khách hàng.
* Customer Since: Ngày khách hàng bắt đầu sử dụng dịch vụ.
* SSN: Mã số an sinh xã hội.
* Phone No.: Số điện thoại của khách hàng.

**Thông tin địa lý**:

* Place Name: Tên địa điểm.
* County: Quận/Huyện.
* City: Thành phố.
* State: Bang/Quốc gia.
* Zip: Mã bưu chính.
* Region: Khu vực.

**Thông tin giao dịch**:

* category: Danh mục sản phẩm.
* payment\_method: Phương thức thanh toán.
* ref\_num: Số tham chiếu giao dịch.
* User Name: Tên người dùng thực hiện giao dịch.
* Discount\_Percent: Phần trăm giảm giá.

### Các thông số thống kê cơ bản

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Thuộc tính** | **Số giá trị duy nhất** | **Giá trị phổ biến nhất (Top)** | **Tần suất (Freq)** | **Trung bình (Mean)** | **Tối thiểu - Tối đa (Min - Max)** |
| order\_id | 201,716 | 100476608 | 43 | N/A | N/A |
| order\_date | 365 | 2020-12-20 | 13,678 | N/A | N/A |
| qty\_ordered | N/A | N/A | N/A | 3.01 | 1 - 501 |
| price | N/A | N/A | N/A | 851.39 | 0.0 - 101,262.59 |
| total | N/A | N/A | N/A | 815.84 | 0.0 - 101,262.59 |
| Discount\_Percent | N/A | N/A | N/A | 6.07 | 0 - 75 |

### Đánh giá tính chất dữ liệu

**Dữ liệu định danh:**

* Các cột như order\_id, item\_id, sku, và cust\_id được sử dụng để định danh duy nhất các đơn hàng, sản phẩm, và khách hàng.

**Dữ liệu thời gian:**

* order\_date cung cấp mốc thời gian cho từng giao dịch, hỗ trợ việc phân tích xu hướng theo ngày, tháng, và năm.

**Dữ liệu số:**

* Các cột như qty\_ordered, price, value, và discount\_amount là các thông tin định lượng cần thiết để tính toán và phân tích doanh thu, lợi nhuận.

**Dữ liệu phân loại:**

* Các cột như payment\_method, category và Region cung cấp thông tin phân nhóm quan trọng để phân tích đặc điểm giao dịch và thị trường.

## Các đặc tính của dữ liệu

Dữ liệu là tập hợp các thông tin có tính tổ chức, và các đặc tính của dữ liệu giúp mô tả chi tiết hơn về tính chất và vai trò của nó trong phân tích.

### Đặc tính về loại dữ liệu

**Dữ liệu định danh (Nominal Data):**

* Là các giá trị không có thứ tự hoặc tính toán, thường dùng để định danh hoặc phân nhóm.
* Ví dụ: order\_id, cust\_id, sku, Region, City, State.

**Dữ liệu thứ tự (Ordinal Data):**

* Là dữ liệu có tính thứ tự, nhưng khoảng cách giữa các giá trị không đồng nhất.

**Dữ liệu định lượng (Quantitative Data):**

* Dữ liệu có thể đo lường và tính toán.
* Bao gồm:
  + Dữ liệu liên tục: price, total, value, Discount\_Percent.
  + Dữ liệu rời rạc: qty\_ordered (số lượng sản phẩm).

**Dữ liệu thời gian (Time Series Data):**

* Là dữ liệu có yếu tố thời gian, thường sử dụng để phân tích xu hướng.
* Ví dụ: order\_date, Customer Since.

### Đặc tính về cấu trúc dữ liệu

**Dữ liệu có cấu trúc (Structured Data):**

* Dữ liệu có định dạng rõ ràng với các cột và hàng, dễ dàng phân tích.
* Toàn bộ bộ dữ liệu thuộc dạng dữ liệu có cấu trúc, phù hợp để xử lý bằng các công cụ phân tích như SQL, Python, R.

**Dữ liệu dạng bảng (Tabular Data):**

* Dữ liệu được tổ chức dưới dạng bảng, trong đó:
  + Dòng: Đại diện cho từng mục chi tiết (chi tiết đơn hàng).
  + Cột: Đại diện cho các thuộc tính như thông tin khách hàng, sản phẩm, địa lý.

### Đặc tính về giá trị dữ liệu

**Tính duy nhất (Uniqueness):**

* Một số cột có giá trị duy nhất, chẳng hạn order\_id hoặc cust\_id.

**Giá trị trùng lặp (Duplicated Values):**

* Một số giá trị như order\_date hoặc Region có thể lặp lại, thể hiện các giao dịch diễn ra trong cùng ngày hoặc khu vực.

**Giá trị ngoại lệ (Outliers):**

* Các giá trị có khoảng cách lớn so với giá trị trung bình. Ví dụ:
  + price có giá trị cao nhất là 101,262.59, lớn hơn nhiều so với giá trị trung bình là 851.39.

**Giá trị thiếu (Missing Values):**

* Dữ liệu có thể chứa các giá trị trống ở một số cột, cần được kiểm tra để xử lý.

### Đặc tính về phân phối dữ liệu

**Phân phối đều (Uniform Distribution):**

* Một số thuộc tính có phân phối đồng đều, như Region, do sự chia đều khu vực địa lý.

**Phân phối không đều (Skewed Distribution):**

* Một số cột như price hoặc qty\_ordered có phân phối lệch với các giá trị cao tập trung ở một số ít dòng.

### Đặc tính về mối quan hệ dữ liệu

**Liên kết giữa các cột:**

* Có các mối quan hệ logic giữa các cột, ví dụ:
  + value = qty\_ordered × price.
  + total = value - discount\_amount.

**Phân cấp dữ liệu:**

* Dữ liệu có cấu trúc phân cấp, như một order\_id có thể chứa nhiều sản phẩm (item\_id).

### Đặc tính về tính nhạy cảm

**Dữ liệu nhạy cảm (Sensitive Data):**

* Một số cột chứa thông tin cá nhân cần bảo mật, như SSN, Phone No., E Mail.

**Dữ liệu bảo mật (Confidential Data):**

* Thông tin tài chính như total hoặc Discount\_Percent cần được bảo vệ để tránh lộ thông tin chiến lược kinh doanh.

### Đặc tính về tính động

Dữ liệu có tính thời gian (Time-Sensitive), có thể thay đổi qua từng năm hoặc từng giai đoạn, đặc biệt ở các cột như order\_date.

### Đặc tính về khả năng sử dụng

**Khả năng khai thác (Exploitable):**

* Dữ liệu được tổ chức rõ ràng, có thể dễ dàng phân tích bằng các công cụ như Power BI, Tableau, Python.

**Khả năng dự đoán (Predictive):**

* Bộ dữ liệu giàu thông tin, hỗ trợ xây dựng các mô hình dự đoán doanh thu, hành vi khách hàng.

# TIỀN XỬ LÝ DỮ LIỆU

## Dữ liệu ban đầu

A group of people in different colors

Description automatically generated

Dữ liệu ban đầu trước khi xử lý bao gồm 35 cột và 286392 dòng

## Xóa các cột dữ liệu không cần thiết

A computer code with text

Description automatically generated with medium confidence

Xóa các cột mang giá trị không liên quan đến bài toán như 'sku', 'bi\_st', 'ref\_num', 'Middle Initial', 'Zip'. Vì những cột này mang tính định danh không ảnh hưởng đến mô hình.

## Kiểm tra xử lý dữ liệu bị thiếu (Missing values)

**Mục tiêu:** Xác định và xử lý các giá trị thiếu trong dataset để tránh ảnh hưởng đến phân tích.

**Cách thực hiện:**

* Kiểm tra giá trị thiếu:

A close-up of a white background

Description automatically generated

Kiểm tra thấy có 1574 giá trị null ở cột **Discount\_Percent**

* Xử lý giá trị thiếu:
  + Đặt các giá trị trống của 'value' bằng giá trị của ‘qty\_ordered’ nhân với ‘price’
  + Đặt các giá trị trống của 'total' bằng ‘value’ trừ đi ‘discount\_amount’
  + Discount\_Percent có 1574 giá trị null nên sẽ xóa đi những dòng có giá trị null này.
* Hướng giải quyết:

A screenshot of a computer

Description automatically generated

Sau khi kiểm tra các dòng có giá trị null, nhận thấy rằng trong đơn hàng có sản phẩm (item\_id) nhưng lại có giá của sản phẩm là 0 nên xóa những dòng này.

## Xử lý kiểu dữ liệu

A screenshot of a computer program

Description automatically generated

Cột qty\_ordered, age có dạng float nên cần chuyển thành kiểu int.

## Xử lý giá trị trùng lặp (Duplicates)

**Mục tiêu:** Loại bỏ các bản ghi trùng lặp để tránh ảnh hưởng đến kết quả phân tích.

**Cách thực hiện:**

* Kiểm tra dữ liệu trùng lặp
* Loại bỏ dữ liệu trùng lặp

A close-up of a computer screen

Description automatically generated

## Kết quả

Dữ liệu còn 214818 dòng và 30 cột

# TRỰC QUAN HÓA DỮ LIỆU

## Doanh thu theo thời gian

### Tổng quan xu hướng

Biểu đồ thể hiện sự biến động của tổng doanh thu theo từng tháng trong giai đoạn từ tháng 10/2020 đến tháng 9/2021.

Doanh thu không ổn định, có những thời điểm tăng đột biến nhưng cũng có những giai đoạn giảm mạnh.

### Các điểm nổi bật

**Tháng 11/2020:** Doanh thu đạt mức cao nhất trong giai đoạn này, khoảng 6×10⁷. Đây là giai đoạn tăng trưởng đột biến so với tháng trước.

**Tháng 12/2020 - Tháng 2/2021:** Doanh thu giảm mạnh và chạm mức thấp nhất vào tháng 2/2021.

**Tháng 3/2021 - Tháng 4/2021:** Doanh thu phục hồi mạnh mẽ, đạt đỉnh thứ hai trong tháng 4/2021.

**Tháng 5/2021:** Sau đỉnh vào tháng 4, doanh thu giảm sâu, đánh dấu một sự sụt giảm lớn.

**Tháng 6/2021 - Tháng 8/2021:** Doanh thu có xu hướng phục hồi nhẹ nhưng không bền vững, duy trì ở mức trung bình.

**Tháng 9/2021:** Doanh thu giảm mạnh trở lại, trở về gần mức thấp.

### Đặc điểm chu kỳ

Có sự tăng trưởng và sụt giảm rõ rệt theo chu kỳ, đặc biệt là vào các tháng cuối năm và đầu năm.

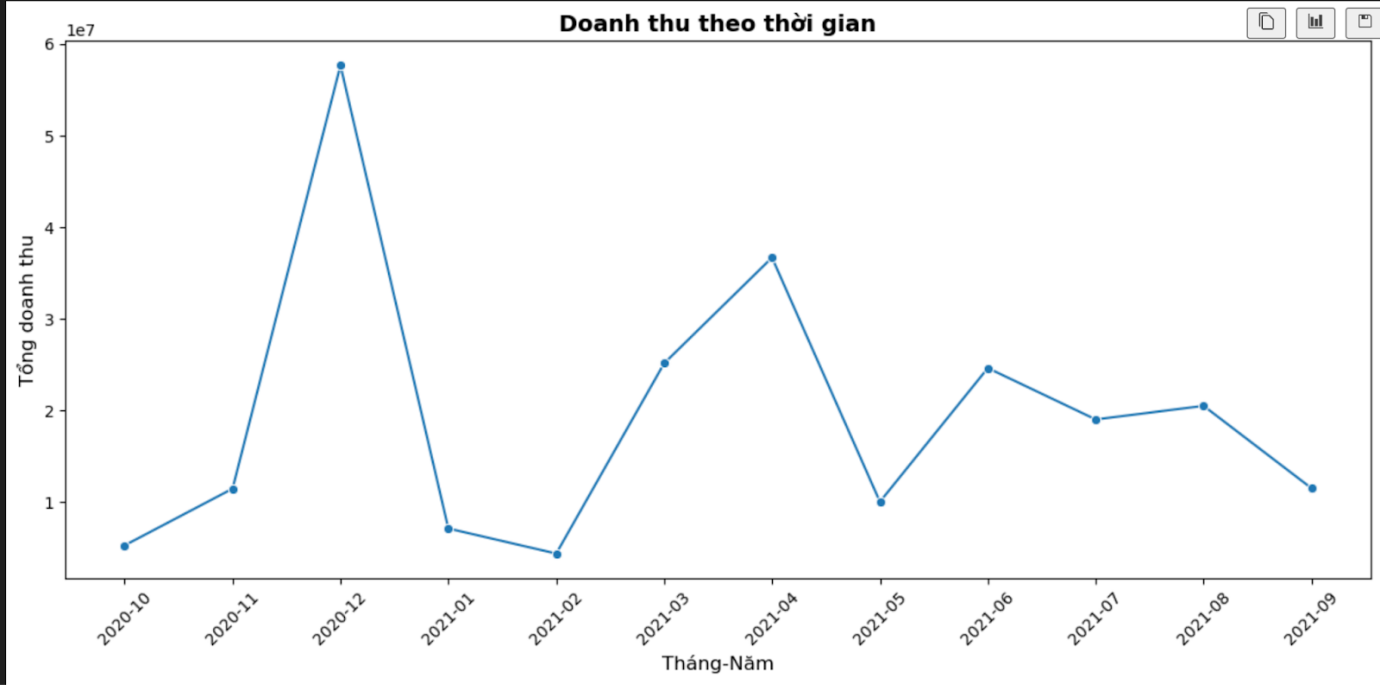
Doanh thu có xu hướng đạt đỉnh vào cuối năm và giảm mạnh vào đầu năm, điều này có thể liên quan đến các yếu tố mùa vụ hoặc các hoạt động kinh doanh cụ thể.

### Kết luận

Biểu đồ cho thấy sự biến động lớn trong tổng doanh thu theo thời gian.

Cần phân tích thêm các yếu tố bên ngoài như mùa vụ, chính sách kinh doanh, hoặc tác động của thị trường để giải thích rõ hơn các đỉnh và đáy doanh thu.

Để duy trì doanh thu ổn định, cần tập trung vào các chiến lược để giảm thiểu sự sụt giảm sau các giai đoạn tăng trưởng cao.



Hình .: Doanh thu theo thời gian

## Phân phối độ tuổi khách hàng

### Tổng quan

Biểu đồ thể hiện tần suất khách hàng theo từng nhóm tuổi, với đường cong biểu diễn xu hướng phân phối.

Độ tuổi khách hàng trải rộng từ 20 đến 70 tuổi, với sự phân bố tương đối đồng đều nhưng có một số điểm dao động rõ rệt.

### Chi tiết phân phối

**Nhóm tuổi từ 20 đến 30:** Tần suất khách hàng ở mức trung bình, dao động quanh mức 10.000.

**Nhóm tuổi từ 30 đến 40:** Khách hàng trong nhóm này khá ổn định, với tần suất cao hơn một chút so với nhóm 20-30.

**Nhóm tuổi từ 40 đến 50:** Có sự sụt giảm mạnh ở giữa độ tuổi 40-50, tần suất thấp hơn đáng kể so với các nhóm khác.

**Nhóm tuổi từ 50 đến 60:** Tần suất tăng trở lại, đạt mức cao nhất ở một số khoảng, vượt 12.000, cho thấy đây là nhóm khách hàng đông đảo nhất.

**Nhóm tuổi từ 60 đến 70:** Tần suất giảm dần về cuối biểu đồ, đặc biệt từ khoảng 65 tuổi trở đi.

### Đặc điểm nổi bật

**Dao động trong phân phối:** Trong khoảng 40-50 tuổi và 60-70 tuổi, có sự dao động lớn về số lượng khách hàng.

**Nhóm khách hàng chủ yếu:** Nhóm tuổi từ 50-60 có tần suất cao nhất, cho thấy đây là độ tuổi đóng vai trò quan trọng trong cơ sở khách hàng.

**Giảm mạnh ở một số khoảng:** Tần suất khách hàng giảm rõ rệt ở độ tuổi 40-50 và trên 65 tuổi, có thể do đặc điểm nhân khẩu học hoặc tính chất sản phẩm/dịch vụ không phù hợp với độ tuổi này.

### Kết luận

Khách hàng tập trung nhiều nhất trong độ tuổi 50-60, phù hợp để tập trung các chiến lược marketing cho nhóm tuổi này.

Cần tìm hiểu nguyên nhân giảm mạnh ở độ tuổi 40-50 và khách hàng trên 65 tuổi để có các biện pháp cải thiện hoặc mở rộng đối tượng khách hàng.

Để tăng cường hiệu quả kinh doanh, nên xây dựng các chiến lược tiếp cận riêng cho từng nhóm tuổi dựa trên đặc điểm phân phối này.



Hình .: Phân phối độ tuổi khách hàng

## Doanh thu theo mục sản phẩm

### Tổng quan

Biểu đồ thể hiện doanh thu phân theo từng danh mục sản phẩm.

Các danh mục sản phẩm có sự chênh lệch lớn về doanh thu, với một số danh mục chiếm tỷ trọng vượt trội so với các danh mục khác.

### Chi tiết từng danh mục

**Mobiles & Tablets:** Đây là danh mục có doanh thu cao nhất, vượt 1.2×10⁸, chiếm phần lớn tổng doanh thu. Điều này cho thấy sản phẩm công nghệ, đặc biệt là điện thoại và máy tính bảng, là mặt hàng được tiêu thụ mạnh nhất.

**Appliances (Thiết bị gia dụng):** Xếp thứ hai với doanh thu khoảng 0.6×10⁸, cho thấy đây là nhóm sản phẩm có sức tiêu thụ lớn, đứng sau sản phẩm công nghệ.

**Entertainment (Giải trí):** Đứng thứ ba với doanh thu gần 0.4×10⁸, phản ánh nhu cầu cao đối với các sản phẩm giải trí như TV, thiết bị âm thanh hoặc game.

**Others (Các sản phẩm khác):** Có doanh thu tương đối lớn, khoảng 0.3×10⁸, cho thấy sự đa dạng của các sản phẩm ngoài các danh mục cụ thể.

**Computing (Máy tính và phụ kiện):** Doanh thu xếp thứ năm, nhưng thấp hơn đáng kể so với các danh mục đứng đầu.

**Thời trang (Women's Fashion & Men's Fashion):** Doanh thu thấp hơn nhiều so với các nhóm công nghệ và gia dụng, cho thấy thời trang không phải là thế mạnh của doanh nghiệp.

Các danh mục khác (Beauty & Grooming, Home & Living, Health & Sports, Kids & Baby, School & Education, Books):

Doanh thu rất thấp và không đáng kể, cho thấy nhu cầu thấp hoặc doanh nghiệp chưa đầu tư mạnh vào các sản phẩm thuộc nhóm này.

### Những điểm nổi bật

**Sự chênh lệch lớn:** Doanh thu của danh mục "Mobiles & Tablets" vượt trội so với các danh mục khác, cho thấy doanh nghiệp đang phụ thuộc lớn vào mặt hàng công nghệ.

**Thế mạnh ở các sản phẩm công nghệ và gia dụng:** Nhóm sản phẩm công nghệ (Mobiles & Tablets, Appliances) chiếm phần lớn tổng doanh thu, phản ánh nhu cầu thị trường cao hoặc chiến lược kinh doanh tập trung vào các mặt hàng này.

**Các danh mục yếu:** Các danh mục như thời trang, sách, sản phẩm dành cho trẻ em và thể thao có doanh thu rất thấp. Đây có thể là những lĩnh vực cần được đầu tư thêm nếu muốn mở rộng thị trường.

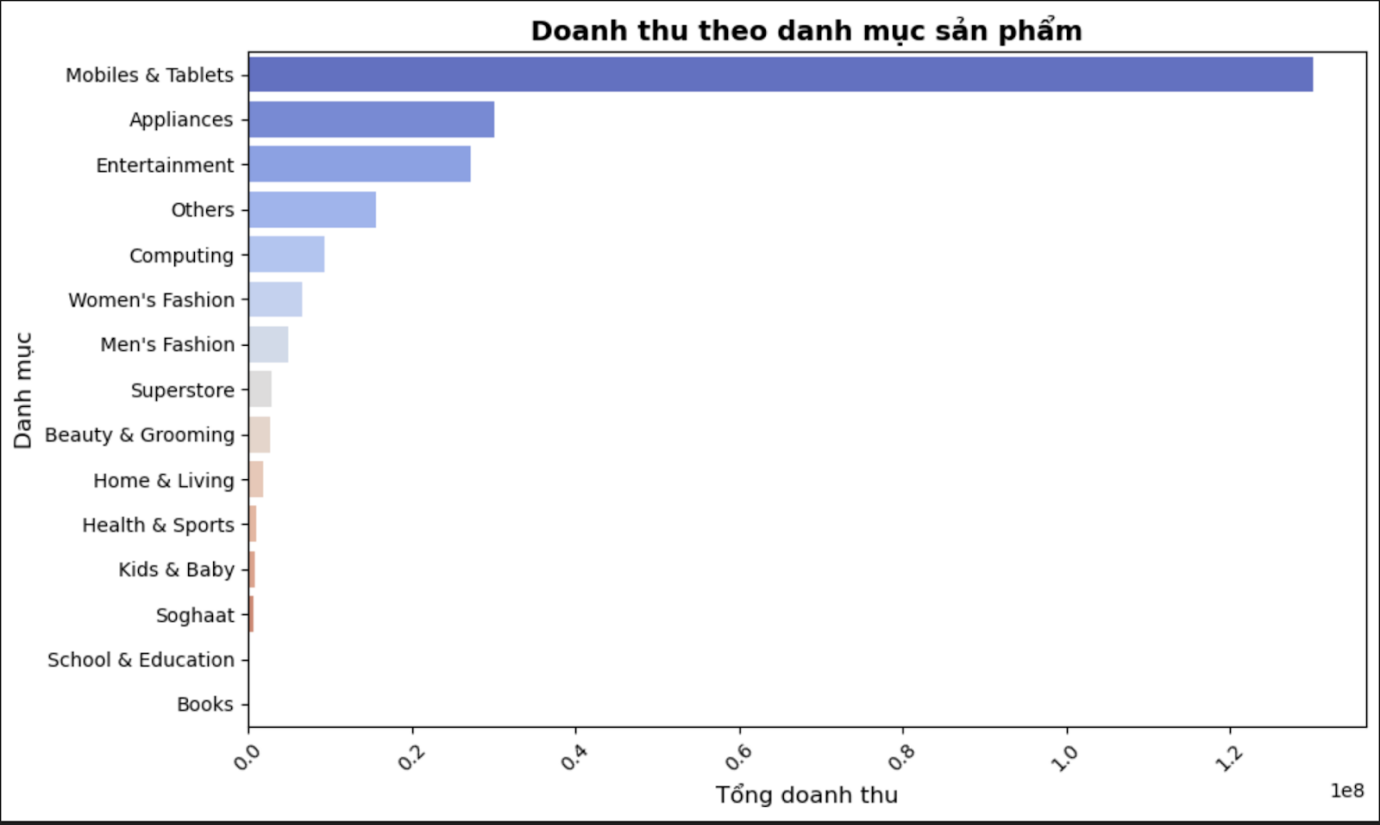
### Kết luận

**Thế mạnh:** Doanh nghiệp đang tập trung và thành công trong việc kinh doanh các sản phẩm công nghệ và gia dụng, đặc biệt là "Mobiles & Tablets".

**Hạn chế:** Các danh mục sản phẩm khác như thời trang, sách, và các sản phẩm dành cho gia đình hoặc trẻ em có doanh thu rất thấp, cho thấy sự mất cân đối trong danh mục sản phẩm.

**Đề xuất:** Duy trì và phát huy thế mạnh ở các danh mục công nghệ và gia dụng.

Xem xét mở rộng hoặc cải thiện chiến lược kinh doanh đối với các danh mục có doanh thu thấp để đa dạng hóa nguồn thu và giảm sự phụ thuộc vào một vài danh mục chính.



Hình .: Tương quan Số lượng - Giảm giá

## Tỷ lệ đơn hàng theo phương thức thanh toán

### Phương thức thanh toán phổ biến nhất

COD (Cash on Delivery) chiếm tỷ lệ cao nhất, với 35.9%. Điều này cho thấy khách hàng vẫn ưa chuộng thanh toán khi nhận hàng, có thể do họ muốn đảm bảo nhận đúng sản phẩm hoặc tránh rủi ro thanh toán online.

### Phương thức thanh toán phổ biến thứ hai

Easypay đứng thứ hai với 24.3%, cho thấy khách hàng bắt đầu quen thuộc với các giải pháp thanh toán kỹ thuật số.

### Các phương thức khác

Payaxis (10.8%) và Easypay\_voucher (10.4%) cũng có tỷ lệ tương đối cao, cho thấy sự đa dạng trong lựa chọn thanh toán kỹ thuật số.

Bankalfalah (8.1%) và một số phương thức khác có tỷ lệ thấp hơn (dưới 5%), có thể là do chúng phục vụ một nhóm đối tượng khách hàng cụ thể.

### Phương thức ít phổ biến nhất

Các phương thức như jazzwallet, Easypay\_MA, và một số khác có tỷ lệ không đáng kể, cho thấy mức độ sử dụng rất thấp. Điều này có thể là do ít người biết đến hoặc không tiện lợi.

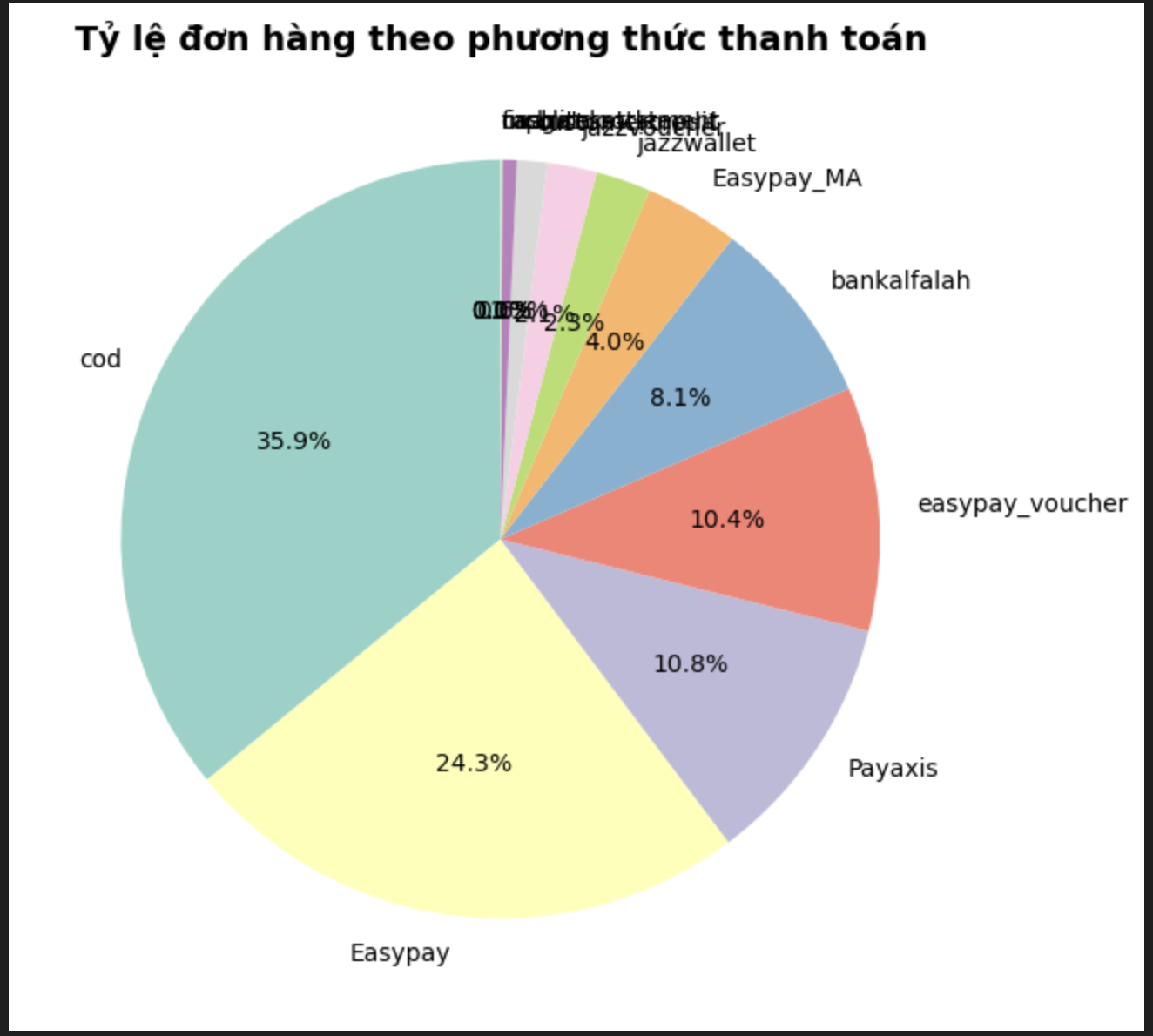
### Nhận định tổng quan

Dù thanh toán qua COD vẫn chiếm tỷ lệ lớn nhất, nhưng các giải pháp thanh toán kỹ thuật số như Easypay và Payaxis đang có xu hướng tăng trưởng, chứng tỏ sự thay đổi hành vi của khách hàng theo hướng tiện lợi và hiện đại hơn. Các phương thức thanh toán ít phổ biến cần được xem xét về tính hiệu quả hoặc cần các chiến lược quảng bá mạnh mẽ hơn.

### Đề xuất

Khuyến khích khách hàng chuyển sang sử dụng các phương thức thanh toán trực tuyến an toàn và tiện lợi.

Phân tích thêm nguyên nhân tại sao các phương thức thanh toán ít phổ biến chưa được sử dụng nhiều và triển khai chiến lược cải thiện.



Hình .: Tỷ lệ đơn hàng theo phương thức thanh toán

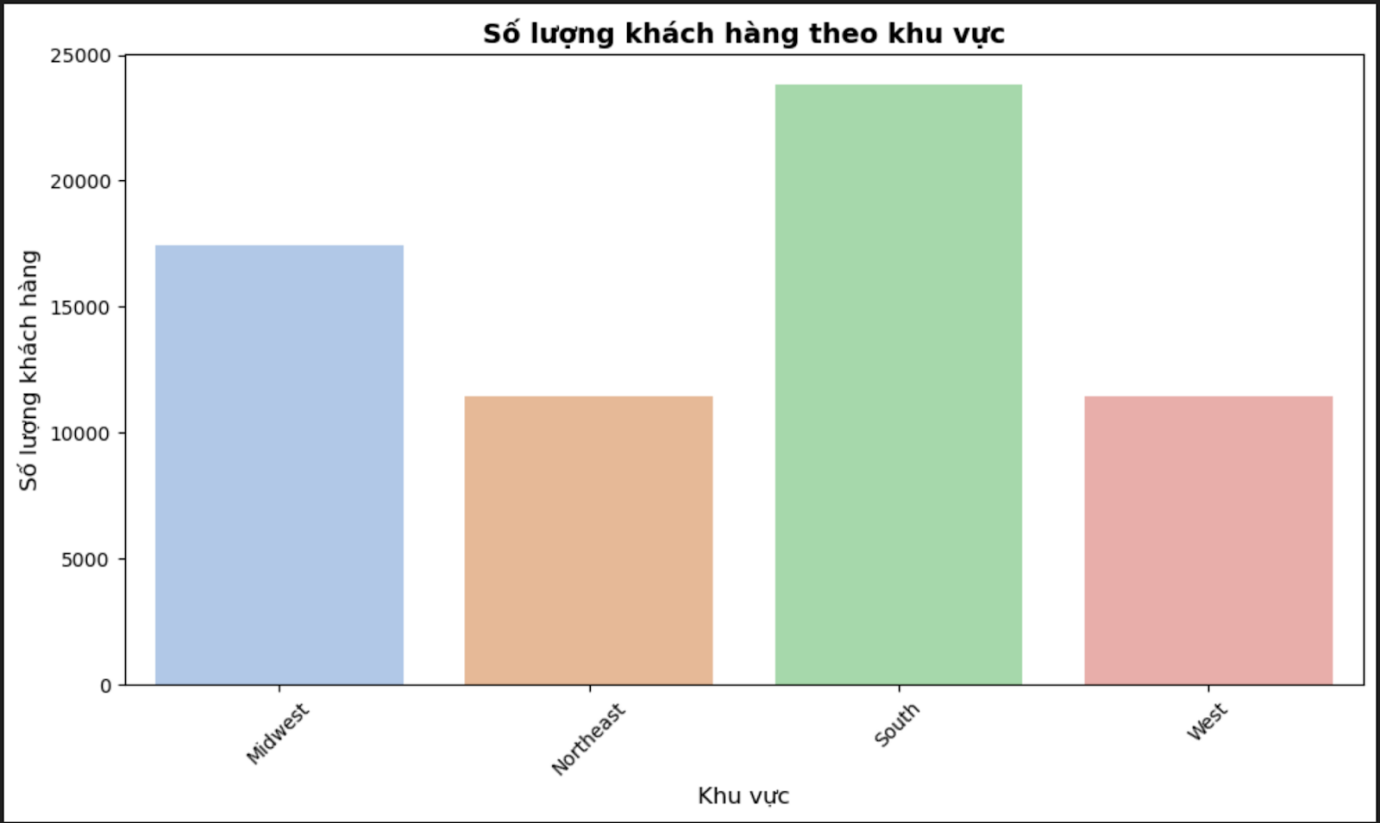
## Số lượng khách hàng theo khu vực

**South** có số lượng khách hàng lớn nhất, khoảng **25,000,** cho thấy đây là thị trường trọng điểm hoặc khu vực này có mức độ quan tâm cao đối với sản phẩm/dịch vụ.

Khu vực có số lượng khách hàng thấp nhất: **Northeast** có số lượng khách hàng thấp nhất, khoảng **10,000**. Điều này có thể liên quan đến chiến lược tiếp cận hoặc nhu cầu thấp hơn tại khu vực này.

So sánh giữa các khu vực: **Midwest** và **West** có số lượng khách hàng tương đối cân bằng, khoảng từ **15,000–20,000**.

Sự khác biệt rõ ràng giữa khu vực dẫn đầu (**South**) và khu vực thấp nhất (**Northeast**) cho thấy sự phân bố không đồng đều về nhu cầu hoặc khả năng tiếp cận khách hàng.



Hình .: Số lượng khách hàng theo khu vực

## Tổng doanh thu theo khu vực

### Khu vực có doanh thu cao nhất

South tiếp tục dẫn đầu với doanh thu cao nhất, trên 10 triệu. Điều này phù hợp với số lượng khách hàng lớn ở khu vực này, đồng thời cho thấy mức chi tiêu cao hơn từ nhóm khách hàng tại đây.

### Khu vực có doanh thu thấp nhất

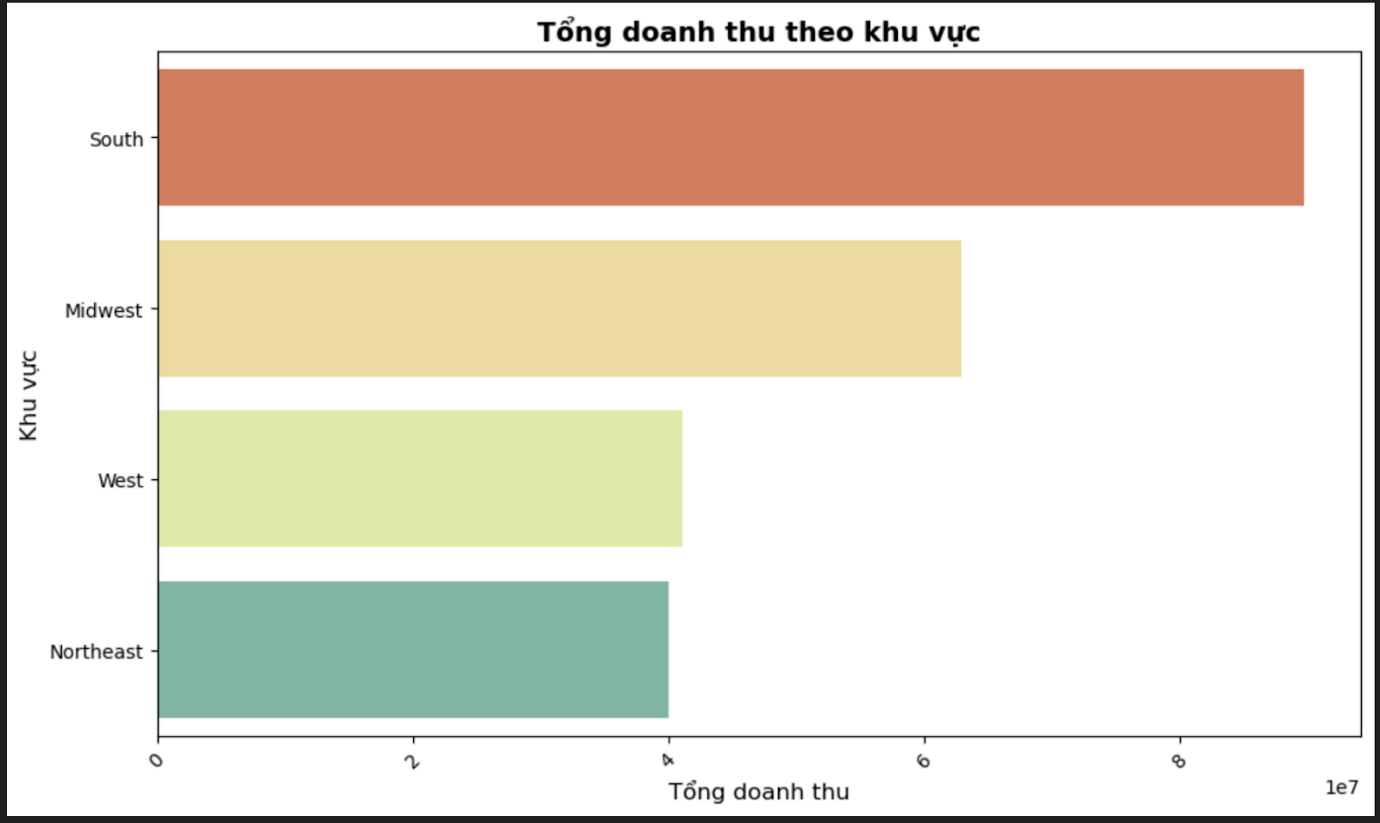
Northeast cũng là khu vực có doanh thu thấp nhất, chưa đạt 5 triệu, phản ánh sự hạn chế về số lượng khách hàng và mức tiêu dùng.

### Tương quan giữa các khu vực

Midwest và West có doanh thu gần bằng nhau, trong khoảng từ 6–8 triệu, cho thấy mức tiêu dùng của khách hàng tại đây không chênh lệch nhiều, mặc dù số lượng khách hàng khác nhau.

### Nhận định tổng quan

Doanh thu tại các khu vực có sự liên quan chặt chẽ với số lượng khách hàng. Tuy nhiên, khu vực South không chỉ dẫn đầu về số lượng khách mà còn có mức tiêu dùng bình quân cao, trong khi Northeast cần cải thiện cả hai yếu tố để tăng trưởng.

****

Hình .: Tổng doanh thu theo khu vực

## Tỉ lệ giảm giá trung bình theo danh mục

### Danh mục có tỷ lệ giảm giá cao nhất

Danh mục Entertainment dẫn đầu với tỷ lệ giảm giá trung bình cao nhất (~10%), theo sau là các danh mục Superstore, Appliances, và Computing. Điều này có thể cho thấy các danh mục này thường xuyên sử dụng chiến lược khuyến mãi để thúc đẩy doanh số.

### Danh mục có tỷ lệ giảm giá thấp

Các danh mục như Books, Others, và School & Education có tỷ lệ giảm giá trung bình thấp nhất. Điều này có thể phản ánh mức độ khuyến mãi hạn chế hoặc sản phẩm ít cạnh tranh hơn.

### Sự chênh lệch trong tỷ lệ giảm giá

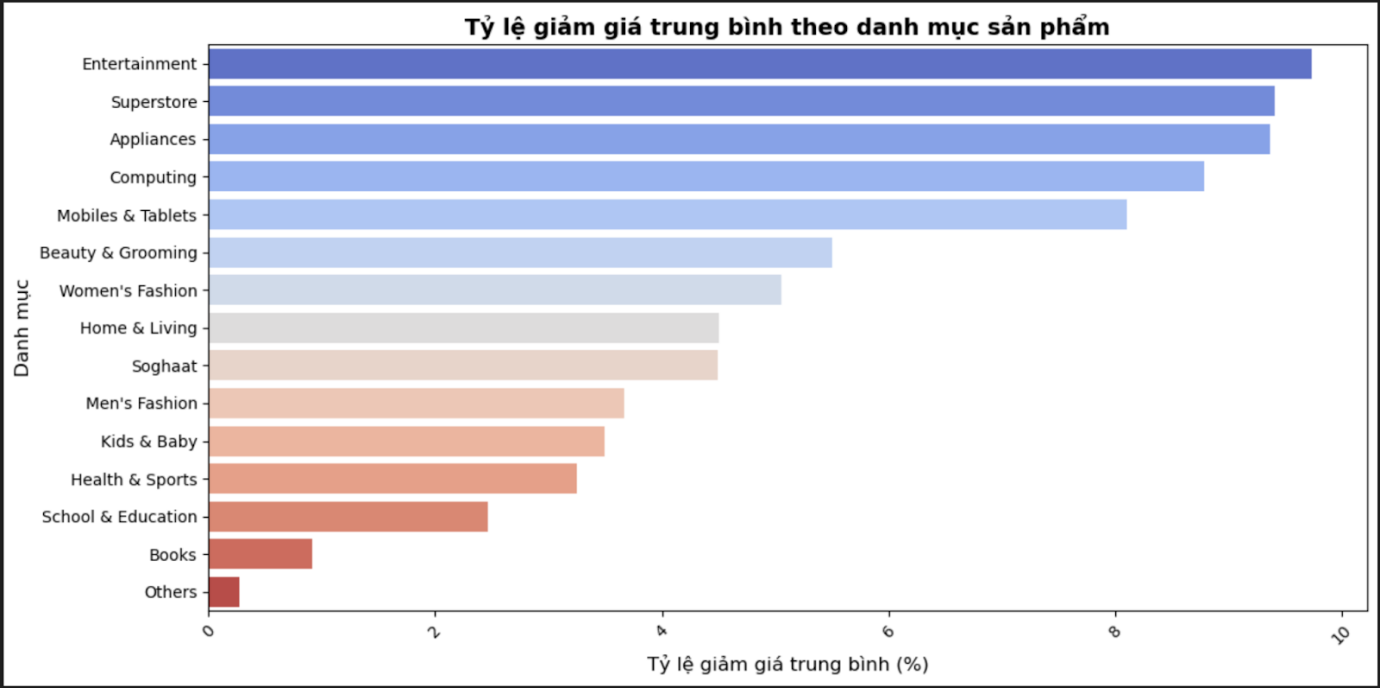
Biểu đồ cho thấy sự chênh lệch lớn giữa các danh mục với nhau. Ví dụ: các danh mục thuộc nhóm giải trí và công nghệ có tỷ lệ giảm giá cao hơn so với nhóm giáo dục và sách.

### Ý nghĩa trong kinh doanh

Tối ưu hóa chiến lược khuyến mãi: Các danh mục có tỷ lệ giảm giá cao có thể được tối ưu để giữ chân khách hàng trung thành và gia tăng doanh thu.

Đánh giá hiệu quả giảm giá: Các danh mục có tỷ lệ giảm giá thấp cần được phân tích xem có nên tăng khuyến mãi để cạnh tranh hoặc giữ mức hiện tại để tối đa hóa lợi nhuận.

Ưu tiên ngân sách marketing: Các danh mục như Entertainment, Superstore nên được ưu tiên đầu tư nếu các chương trình giảm giá đang mang lại doanh số cao.

****

Hình .: Tỉ lệ giảm giá trung bình theo danh mục

# CÔNG NGHỆ VÀ PHƯƠNG PHÁP

## Công nghệ

Hệ thống là một ứng dụng web tích hợp các công cụ và công nghệ hiện đại nhằm quản lý và hiển thị dữ liệu một cách trực quan, hiệu quả. Ứng dụng sử dụng các công nghệ mạnh mẽ như Python, Flask, MongoDB, Chart.js và Tableau để cung cấp các tính năng linh hoạt, dễ sử dụng, đồng thời hỗ trợ xử lý và trực quan hóa dữ liệu.

Sử dụng các mô hình Kmeans, RNN, RandomForest cùng với Chart.js để tương tác dữ liệu trên Web

### Python

**Ngôn ngữ lập trình chính**: Python được chọn vì tính dễ đọc, thư viện phong phú và khả năng tương thích cao với các công cụ và công nghệ khác.

### Chartjs

Thư viện JavaScript được tích hợp trong Flask để trực quan hóa dữ liệu dưới dạng biểu đồ (bar, line, pie chart...).

### Flask

Framework web nhẹ để xây dựng và triển khai ứng dụng. Flask giúp quản lý các luồng xử lý, giao tiếp với cơ sở dữ liệu, và cung cấp các API cho giao diện người dùng.

### TensorFlow

**TensorFlow** là một thư viện mã nguồn mở, mạnh mẽ, do Google phát triển, hỗ trợ xây dựng và triển khai các mô hình học máy (ML) và học sâu (DL). TensorFlow cung cấp các công cụ giúp xử lý dữ liệu, xây dựng mô hình, và tối ưu hóa các thuật toán một cách hiệu quả.

Trong dự án này, TensorFlow được sử dụng để:

* **Xây dựng mô hình học sâu**: Mô hình LSTM (Long Short-Term Memory) trong TensorFlow sẽ được áp dụng để dự đoán số lượng khách hàng hằng ngày dựa trên chuỗi thời gian.
* **Xử lý dữ liệu lớn**: TensorFlow hỗ trợ xử lý khối lượng dữ liệu lớn, tối ưu hóa việc huấn luyện các mô hình học sâu.
* **Triển khai nhanh**: Thư viện TensorFlow cung cấp khả năng triển khai trên nhiều nền tảng, từ máy tính cá nhân đến các hệ thống phân tán.

### Mongodb

 **Cơ sở dữ liệu NoSQL**: MongoDB được sử dụng để lưu trữ dữ liệu dưới dạng tài liệu (JSON-like). Nó phù hợp với hệ thống yêu cầu xử lý và truy vấn dữ liệu phi cấu trúc hoặc bán cấu trúc.

 **Ưu điểm**:

* Linh hoạt trong quản lý dữ liệu.
* Hiệu suất cao khi xử lý lượng lớn dữ liệu.

### Tableau

 **Công cụ trực quan hóa dữ liệu**: Tableau được tích hợp để tạo các dashboard tương tác và chuyên sâu, hỗ trợ người dùng cuối trong việc phân tích và đưa ra quyết định dựa trên dữ liệu.

 **Ứng dụng**:

* Hiển thị dữ liệu bán hàng, khách hàng hoặc bất kỳ loại dữ liệu nào khác một cách trực quan.
* Tạo các biểu đồ phức tạp mà Chart.js không hỗ trợ.

## Phương pháp

### Phân cụm khách hàng

**Phân cụm khách hàng** là một trong những nhiệm vụ quan trọng để hiểu hành vi của khách hàng và xây dựng các chiến lược tiếp cận hiệu quả.

**Giải pháp BI:** Sử dụng thuật toán KMeans để phân cụm khách hàng theo các đặc điểm như độ tuổi, hành vi mua sắm, tần suất mua hàng, v.v. Điều này giúp doanh nghiệp xác định các nhóm khách hàng mục tiêu và đưa ra chiến lược marketing phù hợp.

Phương pháp được áp dụng bao gồm:

* **K-Means**: Phân nhóm khách hàng dựa trên các chỉ số RFM (Recency, Frequency, Monetary), giúp doanh nghiệp xác định các nhóm khách hàng có đặc điểm chung.
* **RFM Analysis**:
  + **Recency**: Khoảng thời gian kể từ lần mua gần nhất.
  + **Frequency**: Tần suất mua hàng.
  + **Monetary**: Tổng giá trị chi tiêu.
* **Ứng dụng K-Means và RFM**: Dựa trên kết quả phân cụm, doanh nghiệp có thể:
  + Tập trung vào nhóm khách hàng trung thành.
  + Triển khai chiến lược tái kích hoạt nhóm khách hàng không còn hoạt động.

### Dự đoán khách hàng quay lại

**Giải pháp BI:** Sử dụng mô hình Random Forest để dự đoán sự quay lại mua hàng của khách hàng. Mô hình này giúp doanh nghiệp xác định các yếu tố ảnh hưởng đến hành vi của khách hàng và đưa ra các biện pháp để giữ chân khách hàng lâu dài.

Các bước thực hiện gồm:

1. **Xử lý dữ liệu**: Chuẩn hóa các biến số, xử lý giá trị thiếu.
2. **Huấn luyện mô hình**: Sử dụng các đặc trưng như Recency, Frequency, Monetary để dự đoán xác suất quay lại của khách hàng.
3. **Đánh giá mô hình**: Sử dụng các chỉ số như Accuracy, F1-Score để đảm bảo chất lượng.

### Dự đoán số lượng khách hàng theo chuỗi thời gian

**Giải pháp BI:** Sử dụng mạng nơ-ron hồi tiếp (RNN) để dự đoán số lượng khách hàng trong tương lai dựa trên các dữ liệu lịch sử. Điều này giúp doanh nghiệp có cái nhìn chính xác về nhu cầu và xu hướng khách hàng trong tương lai, từ đó điều chỉnh chiến lược marketing và phân bổ nguồn lực.

Phương pháp thực hiện:

* **RNN**:
  + **Đặc điểm**: Có khả năng học các mối quan hệ dài hạn trong chuỗi thời gian.
  + **Xử lý dữ liệu đầu vào**: Chuỗi dữ liệu được chuẩn hóa và định dạng thành các mẫu huấn luyện.
  + **Huấn luyện mô hình**: Sử dụng TensorFlow, tối ưu hóa với Adam Optimizer.
  + **Đánh giá kết quả**: Sử dụng các chỉ số MSE, RMSE để đo lường độ chính xác.

## Kết quả

Phân cụm khách hàng thành công, giúp xác định các nhóm mục tiêu cụ thể.

Dự đoán xác suất quay lại của khách hàng với độ chính xác cao.

Dự đoán số lượng khách hàng hằng ngày, hỗ trợ lên kế hoạch kinh doanh.

## Ứng dụng

**Tối ưu chiến lược khuyến mãi:** Nhắm mục tiêu đúng nhóm khách hàng để tăng hiệu quả.

**Dự báo nhu cầu:** Điều chỉnh nguồn lực phù hợp dựa trên dự đoán số lượng khách hàng.

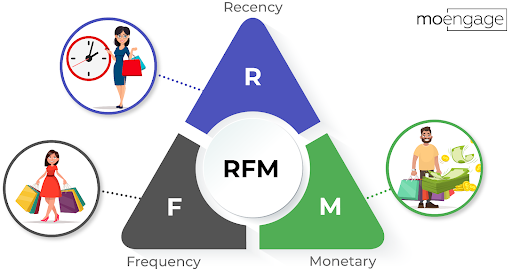
**Cải thiện trải nghiệm khách hàng:** Tập trung vào các nhóm khách hàng trung thành và tiềm năng.

# GIẢI QUYẾT VẤN ĐỀ

## Phân Cụm Khách Hàng theo RFM bằng KMeans

Phân cụm khách hàng là một phương pháp mạnh mẽ để phân loại khách hàng dựa trên các đặc điểm hành vi chung, từ đó giúp doanh nghiệp xây dựng các chiến lược marketing và chăm sóc khách hàng hiệu quả hơn.

Phân cụm theo chỉ số **RFM (Recency, Frequency, Monetary)** kết hợp với thuật toán **KMeans Clustering** là một phương pháp phổ biến trong phân tích hành vi khách hàng, giúp nhóm các khách hàng có hành vi tương tự vào cùng một nhóm (hoặc cụm).



**1. Recency (R) - Mức độ gần đây**

* Khái niệm: Recency đo lường thời gian kể từ lần mua hàng cuối cùng của khách hàng. Khách hàng đã mua hàng gần đây có thể có khả năng mua lại cao hơn.
* Tầm quan trọng: Khách hàng có hành vi mua hàng gần đây (ví dụ: trong vòng 30 ngày) được xem là khách hàng tiềm năng vì họ có xu hướng có nhu cầu cao hơn so với những khách hàng đã lâu không mua.
* Cách tính:
  + Recency có thể được tính bằng số ngày kể từ lần mua hàng cuối cùng của khách hàng.
  + Để phân loại khách hàng, bạn có thể chia Recency thành các khoảng (ví dụ: 0-30 ngày, 31-60 ngày, v.v.).

**2. Frequency (F) - Tần suất mua hàng**

* Khái niệm: Frequency đo lường số lần mà khách hàng đã thực hiện giao dịch trong một khoảng thời gian cụ thể (ví dụ: trong 1 năm, 6 tháng).
* Tầm quan trọng: Khách hàng mua hàng nhiều lần thể hiện rằng họ có sự trung thành và khả năng quay lại mua hàng cao hơn.
* Cách tính:
  + Frequency là tổng số giao dịch của khách hàng trong khoảng thời gian nhất định.
  + Có thể phân loại khách hàng dựa trên tần suất (ví dụ: 1-2 lần mua, 3-5 lần, v.v.).

**3. Monetary (M) - Giá trị chi tiêu**

* Khái niệm: Monetary đo lường tổng số tiền mà khách hàng đã chi tiêu trong một khoảng thời gian nhất định. Khách hàng có giá trị chi tiêu cao thường mang lại nhiều lợi nhuận cho doanh nghiệp.
* Tầm quan trọng: Khách hàng có mức chi tiêu cao thường là những người đem lại doanh thu chính cho doanh nghiệp.
* Cách tính:
  + Monetary là tổng giá trị các giao dịch của khách hàng trong khoảng thời gian (ví dụ: trong 1 năm).
  + Có thể phân loại khách hàng theo các nhóm giá trị chi tiêu (ví dụ: dưới 1 triệu, từ 1 triệu đến 5 triệu, v.v.).

Ba chỉ số này kết hợp lại giúp doanh nghiệp hiểu rõ hơn về hành vi mua sắm của khách hàng, từ đó có thể phân nhóm các khách hàng vào các nhóm có đặc điểm chung.

Chọn các chỉ số **RFM (Recency, Frequency, Monetary)** để phân cụm khách hàng là vì chúng cung cấp cái nhìn rõ ràng về hành vi mua sắm của khách hàng, giúp phân loại khách hàng theo mức độ trung thành, giá trị chi tiêu và thời gian mua hàng gần nhất. **RFM** dễ dàng đo lường từ dữ liệu giao dịch có sẵn, giúp doanh nghiệp hiểu rõ ai là khách hàng tiềm năng, ai trung thành và ai cần tái kích hoạt.

Phương pháp này đơn giản, dễ hiểu và có thể áp dụng rộng rãi cho nhiều loại hình doanh nghiệp, từ bán lẻ đến dịch vụ. Các chỉ số **RFM** cũng hỗ trợ phân tích với các thuật toán phân cụm như **KMeans**, giúp phân nhóm khách hàng hiệu quả, tối ưu chiến lược marketing, và tiết kiệm chi phí chăm sóc khách hàng bằng cách nhắm mục tiêu chính xác.

### Cách Thực Hiện Phân Cụm Khách Hàng theo RFM bằng KMeans

**Bước 1: Chuẩn Bị Dữ Liệu**

Trước tiên, cần thu thập dữ liệu giao dịch của khách hàng, bao gồm các thông tin cần thiết như:

* Ngày giao dịch (để tính Recency).
* Số lần mua hàng (để tính Frequency).
* Tổng số tiền chi tiêu (để tính Monetary).

Dữ liệu này cần được làm sạch, loại bỏ các giá trị thiếu và xử lý các giá trị ngoại lệ để đảm bảo tính chính xác.

**Bước 2: Tính Toán Các Chỉ Số RFM**

Sau khi chuẩn bị xong dữ liệu, tiến hành tính toán ba chỉ số **RFM** cho mỗi khách hàng:

1. **Recency (R):** Tính số ngày kể từ lần mua hàng gần nhất của khách hàng đến thời điểm hiện tại.
2. **Frequency (F):** Tính số lần mua hàng của khách hàng trong khoảng thời gian đã xác định (ví dụ, trong 6 tháng hoặc 1 năm).
3. **Monetary (M):** Tính tổng số tiền mà khách hàng đã chi tiêu trong khoảng thời gian đã chọn.

**Bước 3: Chuẩn Hóa Dữ Liệu**

Vì các chỉ số **RFM** có đơn vị đo khác nhau (ví dụ: **Recency** tính bằng ngày, trong khi **Monetary** tính bằng tiền tệ), cần chuẩn hóa các chỉ số này để đảm bảo các chỉ số có trọng số ngang nhau trong quá trình phân cụm.

Một trong những phương pháp chuẩn hóa phổ biến là **Min-Max Scaling**, giúp đưa tất cả các giá trị về một phạm vi chung, thường là từ 0 đến 1.

Công thức chuẩn hóa **Min-Max Scaling**:

Sau khi chuẩn hóa, các chỉ số **RFM** của mỗi khách hàng sẽ nằm trong phạm vi từ 0 đến 1, giúp thuật toán **KMeans** hoạt động hiệu quả hơn.

**Bước 4: Xác Định Số Lượng Cụm K**

Một trong những bước quan trọng trong phân cụm bằng **KMeans** là xác định số lượng cụm KK. Có thể sử dụng một số phương pháp để xác định số cụm tối ưu, chẳng hạn như:

1. **Elbow Method:** Phương pháp này vẽ đồ thị giữa số cụm KK và tổng sai số bình phương (inertia). Khi số lượng cụm tăng lên, inertia sẽ giảm. Tuy nhiên, có một điểm mà sự giảm này bắt đầu chậm lại, đó là điểm gập khúc, gọi là "elbow". Điểm này giúp xác định số lượng cụm tối ưu.
2. **Silhouette Score:** Là một chỉ số đánh giá chất lượng của phân cụm. Giá trị Silhouette Score dao động từ -1 đến 1, trong đó giá trị gần 1 cho thấy phân cụm tốt. Bạn có thể thử nghiệm với các giá trị KK khác nhau và chọn KK sao cho Silhouette Score đạt giá trị cao nhất.

**Bước 5: Chạy Thuật Toán KMeans**

Khi đã xác định được số lượng cụm k, thuật toán **KMeans** có thể được áp dụng. Các bước chính của thuật toán **KMeans** là:

1. **Khởi tạo tâm cụm**: Thuật toán **KMeans** bắt đầu bằng cách chọn ngẫu nhiên KK điểm làm tâm cụm ban đầu từ dữ liệu. Có thể sử dụng phương pháp **KMeans++** để khởi tạo các tâm cụm sao cho các tâm này phân tán đều hơn, giúp cải thiện chất lượng phân cụm.
2. **Gán điểm dữ liệu vào các cụm**: Sau khi khởi tạo các tâm cụm, thuật toán sẽ tính toán khoảng cách giữa mỗi điểm dữ liệu và các tâm cụm. Mỗi điểm dữ liệu sẽ được gán vào cụm có tâm gần nhất. Khoảng cách giữa điểm dữ liệu xi và tâm cụm μk được tính bằng **khoảng cách Euclid**:
3. **Cập nhật tâm cụm**: Sau khi các điểm dữ liệu đã được phân cụm, tâm của mỗi cụm sẽ được cập nhật bằng trung bình cộng của các điểm trong cụm đó:

Trong đó:

* là tập hợp các điểm dữ liệu trong cụm thứ k
* là số lượng điểm trong cụm thứ k
* là tổng các điểm dữ liệu trong cụm

1. **Lặp lại quá trình**: Các bước phân nhóm và cập nhật tâm cụm sẽ được lặp lại cho đến khi các tâm cụm không thay đổi nữa (thuật toán hội tụ) hoặc khi đạt đến một số vòng lặp tối đa.

**Bước 6: Đánh Giá Kết Quả Phân Cụm**

Sau khi thuật toán **KMeans** hội tụ, có thể đánh giá kết quả phân cụm bằng cách sử dụng các phương pháp đánh giá như:

1. **Silhouette Score**: Đây là một chỉ số đánh giá mức độ phân tách và sự tương đồng của các cụm. Giá trị Silhouette Score gần 1 cho thấy phân cụm hiệu quả.
2. **Elbow Method**: Nếu sử dụng phương pháp **Elbow Method** để xác định số cụm KK, bạn có thể kiểm tra sự thay đổi của inertia và xác nhận rằng điểm gập khúc đã đạt được.

**Bước 7: Trực Quan Hóa Kết Quả Phân Cụm**

Để dễ dàng quan sát các nhóm khách hàng, có thể sử dụng phương pháp **PCA (Principal Component Analysis)** để giảm chiều dữ liệu và trực quan hóa kết quả phân cụm trong không gian 2D hoặc 3D. Điều này giúp dễ dàng nhận diện các nhóm khách hàng và phân tích các đặc điểm chung của từng nhóm.

**Bước 8: Phân Tích và Áp Dụng Kết Quả Phân Cụm**

Sau khi phân cụm thành công, có thể phân tích các nhóm khách hàng:

* **Nhóm khách hàng trung thành(High Value)**: Đây là những khách hàng có **Frequency** và **Monetary** cao. Doanh nghiệp có thể tập trung vào các chương trình khách hàng thân thiết để giữ chân nhóm này.
* **Nhóm khách hàng tiềm năng(Medium Value)**: Đây là những khách hàng có **Recency** cao (mới mua hàng), nhưng **Frequency** và **Monetary** thấp. Doanh nghiệp có thể triển khai các chiến dịch khuyến mãi để kích thích họ mua hàng nhiều hơn.
* **Nhóm khách hàng không hoạt động(Low Value)**: Đây là những khách hàng có **Recency** thấp (không mua hàng trong một thời gian dài). Doanh nghiệp có thể triển khai các chiến lược tái kích hoạt như gửi email, giảm giá đặc biệt để thu hút họ quay lại.

### Lợi Ích Của Phân Cụm Khách Hàng theo RFM

Phân cụm khách hàng bằng KMeans và RFM mang lại nhiều lợi ích cho doanh nghiệp:

**Tối ưu hóa chiến lược marketing:** Các nhóm khách hàng có hành vi tương tự nhau sẽ giúp doanh nghiệp thiết kế các chiến dịch marketing phù hợp, từ đó tăng khả năng thành công.

**Quản lý mối quan hệ khách hàng (CRM):** Việc phân loại khách hàng giúp doanh nghiệp hiểu rõ hơn về nhóm khách hàng mục tiêu và điều chỉnh các chiến lược chăm sóc khách hàng sao cho hiệu quả.

**Tăng trưởng doanh thu:** Tập trung vào các nhóm khách hàng tiềm năng hoặc khách hàng trung thành có thể giúp tăng trưởng doanh thu một cách bền vững.

**Tăng cường dự đoán hành vi khách hàng:** Phân cụm giúp doanh nghiệp dự đoán xu hướng hành vi của khách hàng, từ đó đưa ra các quyết định kịp thời và chính xác.

## Dự Đoán Sự Trở Lại Mua Hàng của Khách Hàng (Random Forest)

### Giới Thiệu về Random Forest

**Random Forest** là một thuật toán học máy mạnh mẽ thuộc nhóm các mô hình cây quyết định. Thuật toán này sử dụng phương pháp **ensemble learning**, trong đó nhiều mô hình (các cây quyết định) được huấn luyện độc lập và sau đó kết hợp kết quả của các mô hình này để đưa ra dự đoán cuối cùng.

Mục tiêu của Random Forest là giảm thiểu sự phức tạp và độ chính xác thấp của các mô hình cây quyết định đơn lẻ bằng cách xây dựng một **rừng cây** với nhiều cây quyết định nhỏ.

**Cây quyết định** là một thuật toán học máy trong đó dữ liệu được chia thành các nhánh (branches) dựa trên các đặc trưng của dữ liệu để đưa ra quyết định cuối cùng. Một cây quyết định có thể được sử dụng để phân loại hoặc hồi quy các giá trị. Trong trường hợp của Random Forest, mỗi cây quyết định sẽ đưa ra một dự đoán (ví dụ: "Rời bỏ" hoặc "Quay lại"), và kết quả cuối cùng của mô hình sẽ là kết quả phổ biến từ tất cả các cây quyết định trong rừng.

Random Forest có thể giải quyết cả bài toán phân loại (như dự đoán sự rời bỏ của khách hàng) và bài toán hồi quy (dự đoán giá trị liên tục như số tiền chi tiêu của khách hàng). Ở đây, chúng ta sẽ áp dụng Random Forest cho bài toán **phân loại** nhằm dự đoán xem liệu khách hàng có khả năng rời bỏ dịch vụ hay quay lại mua hàng hay không.

### Nguyên Lý Hoạt Động của Random Forest

#### **Tạo Rừng Cây Quyết Định**:

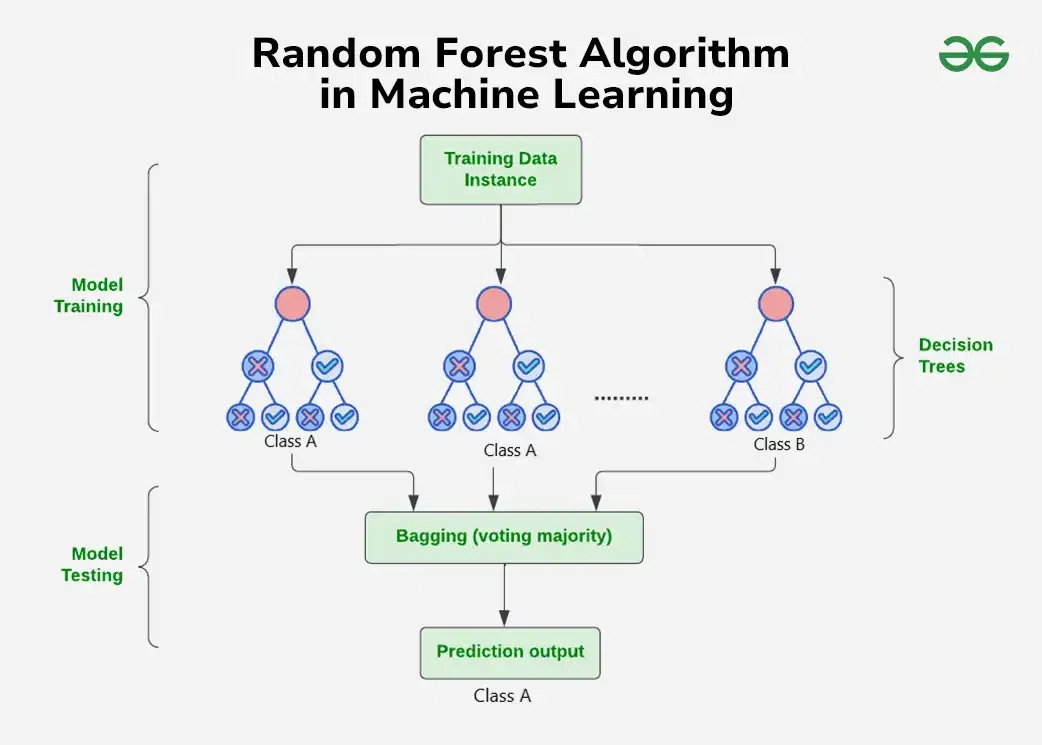
**Bootstrap Aggregating (Bagging)**: Phương pháp này được sử dụng trong Random Forest để xây dựng nhiều cây quyết định từ các mẫu dữ liệu con được lấy mẫu ngẫu nhiên từ tập dữ liệu huấn luyện. Mỗi cây quyết định trong rừng sẽ học từ một phần dữ liệu khác nhau, giúp tăng tính đa dạng và giảm độ lệch của mô hình.

#### **Phân Chia Dữ Liệu**

Mỗi cây quyết định trong rừng sẽ phân chia dữ liệu tại các điểm cắt khác nhau, sao cho mỗi phân đoạn của dữ liệu càng "thuần nhất" càng tốt. Quá trình phân chia này sử dụng các đặc trưng độc lập trong dữ liệu (ví dụ: Recency, Frequency, Monetary, Age) để quyết định xem liệu khách hàng có khả năng rời bỏ hay quay lại.

#### **Dự Đoán**

Sau khi các cây trong rừng được huấn luyện, Random Forest sẽ dựa vào **vote (bỏ phiếu)** của tất cả các cây quyết định để đưa ra kết quả cuối cùng. Mỗi cây quyết định sẽ đưa ra một dự đoán (ví dụ: "Rời bỏ" hoặc "Quay lại"), và kết quả cuối cùng là nhãn mà có số lượng cây dự đoán nhiều nhất. Đây là quá trình kết hợp kết quả từ các cây để giảm thiểu sai số và tăng tính chính xác.



### Các Yếu Tố Ảnh Hưởng Đến Dự Đoán Sự Quay Lại Mua Hàng

Để dự đoán khả năng khách hàng quay lại hoặc rời bỏ, các yếu tố sau có thể ảnh hưởng đến mô hình:

1. **Recency (R)**: Khoảng thời gian kể từ lần mua hàng gần nhất. Khách hàng mua gần đây có khả năng quay lại cao hơn.
2. **Frequency (F)**: Tần suất mua hàng của khách hàng trong một khoảng thời gian. Khách hàng mua hàng thường xuyên có xu hướng trung thành hơn.
3. **Monetary (M)**: Tổng số tiền chi tiêu của khách hàng. Những khách hàng chi tiêu nhiều có thể là nhóm khách hàng quan trọng cần chăm sóc để giữ lại.
4. **Age**: Độ tuổi của khách hàng có thể ảnh hưởng đến quyết định mua sắm và thói quen tiêu dùng.
5. **Discount (Giảm giá)**: Khách hàng nhận giảm giá có thể quay lại để tận dụng các ưu đãi.
6. **Payment Method (Phương thức thanh toán)**: Phương thức thanh toán có thể giúp nhận diện thói quen thanh toán của khách hàng.

### Các Bước Tiền Xử Lý Dữ Liệu

Trước khi áp dụng mô hình Random Forest, cần thực hiện một số bước tiền xử lý dữ liệu:

#### **Chuẩn Hóa Dữ Liệu**:

Các biến như Recency, Frequency, và Monetary cần được chuẩn hóa để các đặc trưng có đơn vị đo khác nhau có thể làm việc cùng nhau. Các phương pháp chuẩn hóa phổ biến là Min-Max Scaling hoặc Standardization.

#### **Mã Hóa Các Biến Phân Loại**:

Các biến như **Gender**, **Payment Method**, **Region** là các biến phân loại cần được mã hóa dưới dạng **One-hot encoding** hoặc **Label encoding**.

#### **Tạo Biến Mới**:

Bạn có thể tạo thêm các biến từ các trường dữ liệu hiện tại, chẳng hạn như **Loyalty Score** (tính từ số lần mua hàng và tổng chi tiêu) hoặc **Customer Age Group** (phân nhóm độ tuổi khách hàng).

#### **Chia Dữ Liệu**:

Dữ liệu cần được chia thành hai phần: **Training Set** (80%) để huấn luyện mô hình và **Test Set** (20%) để đánh giá mô hình.

### Xây Dựng Mô Hình Dự Đoán

Sau khi hoàn tất tiền xử lý, có thể xây dựng mô hình dự đoán sự rời bỏ/quay lại mua hàng của khách hàng bằng Random Forest thông qua các bước sau:

#### **Khởi Tạo Mô Hình**:

Mô hình Random Forest sẽ được khởi tạo với các tham số như số lượng cây quyết định (n\_estimators) và độ sâu tối đa của mỗi cây (max\_depth).

#### **Huấn Luyện Mô Hình**:

Sử dụng tập huấn luyện để huấn luyện mô hình với các biến đặc trưng và nhãn mục tiêu (tức là nhãn "Rời bỏ" hoặc "Quay lại").

#### **Dự Đoán**:

Sau khi huấn luyện, mô hình sẽ sử dụng dữ liệu kiểm tra để dự đoán liệu khách hàng có quay lại mua hàng hay không.

**Đánh Giá Mô Hình**:  
Để đánh giá mức độ chính xác của mô hình, bạn có thể sử dụng các chỉ số như **Accuracy** và **F1-Score**. Các chỉ số này giúp bạn đánh giá khả năng phân loại của mô hình trong việc dự đoán sự rời bỏ hoặc quay lại mua hàng của khách hàng.

* **Accuracy:** Là tỷ lệ dự đoán đúng của mô hình, tính bằng số lượng dự đoán chính xác chia cho tổng số dự đoán.
* **F1-Score:** Là chỉ số kết hợp giữa **Precision** và **Recall**, thường được sử dụng khi bạn muốn có sự cân bằng giữa hai chỉ số này, đặc biệt trong các bài toán phân loại không cân bằng (ví dụ, nhóm khách hàng quay lại ít hơn nhóm rời bỏ).
  + **Precision:** Là tỷ lệ dự đoán đúng các khách hàng quay lại so với tổng số khách hàng mà mô hình dự đoán quay lại.
  + **Recall:** Là tỷ lệ dự đoán đúng các khách hàng quay lại so với tổng số khách hàng thực tế quay lại.

Bạn có thể sử dụng các chỉ số bổ sung như **Confusion Matrix** để đánh giá chất lượng phân loại của mô hình, giúp bạn hiểu rõ hơn về số lượng dự đoán đúng và sai cho từng loại khách hàng (quay lại hoặc rời bỏ).

#### **Tinh Chỉnh Tham Số Mô Hình**

Khi đã đánh giá được hiệu quả mô hình, bạn có thể tiếp tục tối ưu hóa mô hình bằng cách điều chỉnh các tham số của **Random Forest** để cải thiện độ chính xác. Các tham số quan trọng của Random Forest bao gồm:

* **n\_estimators (Số lượng cây quyết định):** Là số lượng cây trong rừng. Tăng số cây có thể giúp giảm độ lệch và cải thiện độ chính xác, nhưng sẽ tăng thời gian huấn luyện.
* **max\_depth (Độ sâu tối đa của mỗi cây):** Giới hạn độ sâu của cây giúp tránh overfitting. Độ sâu quá lớn có thể khiến mô hình học quá chi tiết từ dữ liệu huấn luyện, dẫn đến overfitting.
* **min\_samples\_split (Số mẫu tối thiểu để phân chia một nút):** Quy định số mẫu tối thiểu để phân chia một nút trong cây. Tăng giá trị này có thể giúp giảm độ phức tạp của cây, tránh overfitting.
* **min\_samples\_leaf (Số mẫu tối thiểu tại lá):** Quy định số mẫu tối thiểu ở mỗi lá của cây quyết định. Tham số này giúp kiểm soát độ phức tạp của mô hình và ngăn ngừa overfitting.
* **max\_features (Số lượng đặc trưng tối đa để xét tại mỗi phân chia):** Tham số này giúp giảm sự phụ thuộc vào các đặc trưng mạnh và tăng tính ngẫu nhiên trong mô hình.
* **random\_state (Hạt giống ngẫu nhiên):** Đảm bảo rằng các kết quả huấn luyện mô hình là tái lập được.

**Cách tối ưu tham số:** Bạn có thể sử dụng các phương pháp như **Grid Search** hoặc **Random Search** để tìm ra bộ tham số tốt nhất cho mô hình của mình.

### Trả Về Phần Trăm Quay Lại của Mỗi Khách Hàng

Sau khi huấn luyện mô hình Random Forest,có thể lấy xác suất mà mô hình dự đoán khách hàng sẽ quay lại (label = 1). Xác suất này có thể tính toán bằng cách sử dụng predict\_proba() trong sklearn.

#### Phần Trăm Ảnh Hưởng Của Mỗi Feature Tới Mô Hình

Mô hình Random Forest tính toán **Feature Importance** dựa trên cách mà các đặc trưng ảnh hưởng đến khả năng phân chia các nút trong các cây quyết định. Mỗi đặc trưng sẽ có một giá trị quan trọng, cho thấy mức độ ảnh hưởng của nó đến quyết định phân loại cuối cùng.

Mỗi đặc trưng sẽ có giá trị từ 0 đến 1, và tổng của tất cả các giá trị này sẽ bằng 1. Bạn có thể sử dụng phương pháp feature\_importances\_ trong sklearn để tính toán giá trị này.

#### **Áp Dụng Kết Quả Dự Đoán**

Sau khi huấn luyện và đánh giá mô hình, bạn có thể áp dụng kết quả dự đoán để thực hiện các chiến lược giữ chân khách hàng, bao gồm:

1. **Chiến Lược Marketing Targeted**:
   * Nhắm mục tiêu các khách hàng có khả năng rời bỏ cao bằng các chương trình khuyến mãi, giảm giá hoặc ưu đãi đặc biệt.
2. **Chăm Sóc Khách Hàng**:
   * Tập trung vào nhóm khách hàng có khả năng quay lại và xây dựng chương trình khách hàng trung thành để duy trì mối quan hệ lâu dài.
3. **Cải Thiện Trải Nghiệm Khách Hàng**:
   * Sử dụng dữ liệu từ mô hình dự đoán để cải thiện các yếu tố như dịch vụ khách hàng, tốc độ giao hàng và hỗ trợ kỹ thuật.

Để ứng dụng kết quả dự đoán lên hệ thống BI, có thể hiển thị các thông tin sau:

* **Khách hàng có nguy cơ rời đi cao:** Đây là nhóm khách hàng có khả năng rời bỏ dịch vụ hoặc không quay lại mua hàng cao, dựa trên các yếu tố như thời gian kể từ lần mua hàng gần nhất (Recency), tần suất mua hàng (Frequency), số tiền chi tiêu (Monetary), hoặc các yếu tố khác trong mô hình dự đoán.Việc xác định nhóm khách hàng này giúp doanh nghiệp có thể chủ động áp dụng các chiến lược giữ chân, như gửi email chăm sóc, đưa ra các chương trình khuyến mãi đặc biệt hoặc chăm sóc cá nhân hóa. Đây là những khách hàng quan trọng cần được ưu tiên để giảm thiểu tỉ lệ churn (rời bỏ).
* **Phần trăm khách hàng ở lại:**Tỷ lệ khách hàng có khả năng quay lại mua hàng trong tương lai, dựa trên các yếu tố hành vi lịch sử của họ. Các khách hàng có tỷ lệ quay lại cao có xu hướng trung thành với thương hiệu hoặc dịch vụ.Thể hiện sự ổn định của khách hàng trong việc quay lại mua hàng. Đây là chỉ số quan trọng để đánh giá mức độ thành công của các chiến lược giữ chân khách hàng hiện tại. Tỷ lệ này càng cao, doanh nghiệp càng có thể yên tâm về sự bền vững trong tương lai.
* **Số lượng khách hàng trong quý:** Đây là số lượng khách hàng đã mua hàng trong một quý cụ thể. Thông tin này giúp doanh nghiệp hiểu rõ về sự tham gia và hành vi mua sắm của khách hàng trong một khoảng thời gian nhất định.Thống kê này giúp doanh nghiệp xác định được quy mô và sự phát triển của cơ sở khách hàng trong từng quý. Nó cũng có thể giúp theo dõi sự thay đổi trong hành vi tiêu dùng của khách hàng theo thời gian, chẳng hạn như tăng trưởng hay suy giảm trong số lượng khách hàng.
* **Các đặc trưng ảnh hưởng lớn đến mô hình:** Các đặc trưng (hoặc yếu tố) quan trọng trong dữ liệu mà mô hình Random Forest sử dụng để dự đoán hành vi của khách hàng. Ví dụ: Recency (thời gian gần đây nhất khách hàng mua hàng), Frequency (tần suất mua hàng), Monetary (số tiền chi tiêu), Age (độ tuổi khách hàng) có thể là những yếu tố quyết định.**:** Hiển thị các đặc trưng này giúp doanh nghiệp hiểu rõ hơn về những yếu tố ảnh hưởng đến quyết định của khách hàng, từ đó cải thiện các chiến lược marketing và chăm sóc khách hàng. Ví dụ, nếu "Recency" có ảnh hưởng mạnh mẽ đến việc khách hàng quay lại, doanh nghiệp có thể tăng cường chiến lược gửi thông tin hoặc khuyến mãi cho những khách hàng đã lâu không mua hàng.

### Lợi Ích của Việc Sử Dụng Random Forest

**Giảm Thiểu Overfitting**: Random Forest giảm thiểu overfitting bằng cách xây dựng nhiều cây quyết định từ các tập dữ liệu ngẫu nhiên và sử dụng các đặc trưng ngẫu nhiên tại mỗi phân chia.

**Tăng Độ Chính Xác**: Việc kết hợp nhiều cây quyết định giúp tăng độ chính xác tổng thể của mô hình so với việc chỉ sử dụng một cây quyết định đơn lẻ.

**Khả Năng Xử Lý Dữ Liệu Phức Tạp**: Random Forest có thể xử lý các dữ liệu phức tạp với nhiều đặc trưng và ít bị ảnh hưởng bởi các mối quan hệ phi tuyến tính giữa các đặc trưng.

## Dự Đoán Số Lượng Khách Hàng Theo Ngày (RNN)

Dự đoán số lượng khách hàng có thể đến thăm hoặc mua hàng vào một ngày cụ thể là một bài toán quan trọng đối với các doanh nghiệp, giúp tối ưu hóa việc phân bổ nguồn lực và lên kế hoạch hoạt động. Việc sử dụng các mô hình học sâu, đặc biệt là các mạng nơ-ron hồi tiếp (RNN - Recurrent Neural Networks), giúp dự đoán sự thay đổi trong nhu cầu khách hàng qua thời gian dựa trên dữ liệu lịch sử.

Trong bài toán này, RNN có thể học được các mẫu hành vi của khách hàng từ các chuỗi thời gian, từ đó đưa ra các dự đoán chính xác hơn về số lượng khách hàng trong tương lai.

### Giới Thiệu về RNN

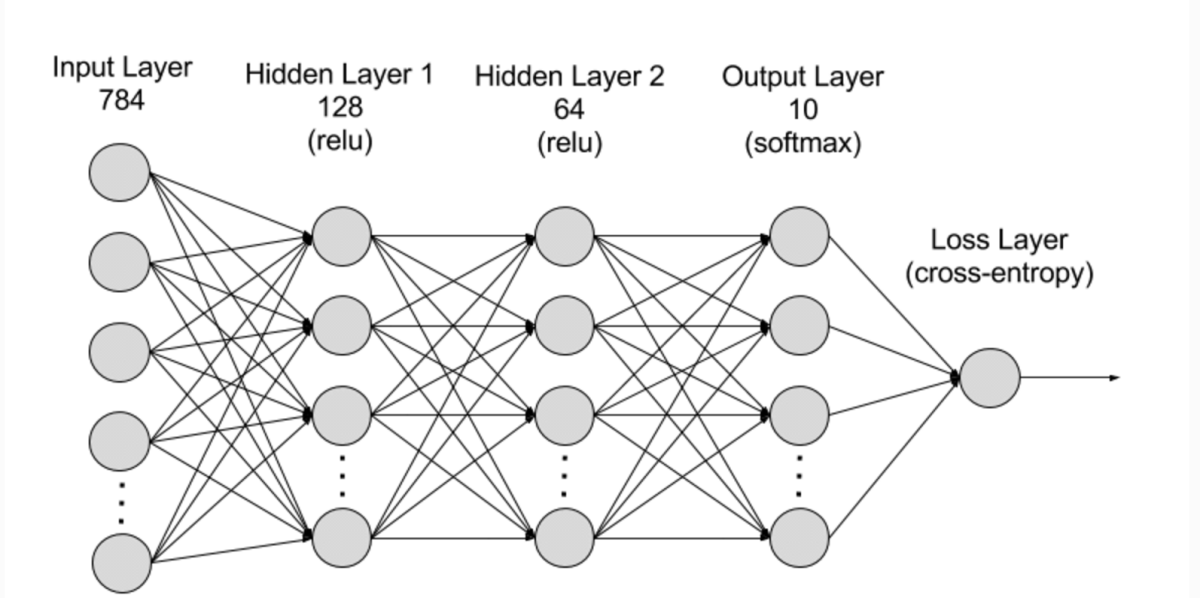
Mạng nơ-ron hồi tiếp (RNN) là một loại mô hình học sâu đặc biệt phù hợp với các bài toán có tính chất chuỗi thời gian (time series), nơi thông tin từ các bước thời gian trước có thể ảnh hưởng đến dự đoán ở bước hiện tại. Khác với các mạng nơ-ron truyền thống, RNN có khả năng lưu trữ thông tin từ các bước thời gian trước và sử dụng chúng trong quá trình tính toán ở các bước tiếp theo. RNN có một cấu trúc đặc biệt với các liên kết vòng (feedback loops), cho phép mô hình duy trì thông tin qua các bước thời gian.

Một biến thể phổ biến của RNN là **LSTM (Long Short-Term Memory)**. LSTM cải thiện khả năng xử lý các vấn đề với độ trễ dài, giúp mô hình tránh hiện tượng **biến mất gradient** — một vấn đề phổ biến trong các RNN cơ bản khi làm việc với các chuỗi thời gian dài. LSTM có các "cổng" (gates) điều khiển thông tin nào được lưu giữ và thông tin nào bị loại bỏ trong quá trình truyền qua các bước thời gian.

### Cấu Trúc và Cách Hoạt Động của RNN

#### Cấu Trúc của RNN

Cấu trúc của một RNN bao gồm ba thành phần chính:

1. **Lớp đầu vào (Input layer):** Đây là nơi mô hình tiếp nhận dữ liệu đầu vào tại mỗi thời điểm t (thường là các yếu tố đầu vào như số lượng khách hàng trong ngày, chương trình khuyến mãi, mùa vụ, ngày lễ, v.v.). Dữ liệu này có thể là một chuỗi giá trị mà mỗi giá trị phụ thuộc vào thời gian trước đó.
2. **Lớp ẩn (Hidden layer):** Lớp này có nhiệm vụ xử lý thông tin nhận được từ lớp đầu vào và lưu trữ các thông tin quan trọng từ các bước thời gian trước đó. Ở mỗi bước thời gian, RNN sử dụng thông tin từ trạng thái ẩn trước đó và dữ liệu đầu vào hiện tại để tính toán trạng thái ẩn mới. Cấu trúc của RNN là hồi tiếp (feedback), nghĩa là trạng thái ẩn sẽ được cập nhật qua các bước thời gian và "truyền" thông tin qua từng bước.
3. **Lớp đầu ra (Output layer):** Đây là nơi mô hình đưa ra kết quả dự đoán tại mỗi thời điểm t. Đối với bài toán dự đoán số lượng khách hàng, lớp đầu ra sẽ đưa ra giá trị dự đoán cho số lượng khách hàng trong ngày tiếp theo (hoặc nhiều ngày trong tương lai).  
   

#### Cách Hoạt Động của RNN

Mỗi bước thời gian trong RNN sẽ tính toán một trạng thái ẩn mới dựa trên dữ liệu đầu vào tại thời điểm đó và trạng thái ẩn của thời điểm trước. Cụ thể, cách hoạt động của RNN có thể mô tả qua các công thức sau:

1. **Cập nhật trạng thái ẩn:**

Trong đó:

* là trạng thái ẩn tại thời điểm t.
* là trạng thái ẩn tại thời điểm trước (t-1).
* là dữ liệu đầu vào tại thời điểm t (Ví dụ: số lượng khách hàng, thông tin về chương trình khuyến mãi,…).
* và là các ma trận trọng số được qua quá trình huấn luyện.
* là bias (hệ số điều chỉnh) cho lớp ẩn.
* Hàm f là một hàm kích hoạt như tanh hoặc ReLU để giới hạn giá trị của trạng thái ẩn.

1. **Dự đoán đầu ra:**

Trong đó:

* là giá trị dự đoán tại thời điểm t (Ví dụ: số lượng khách hàng dự đoán).
* là ma trận trọng số từ trạng thái ẩn đến đầu ra.
* là bias cho lớp đầu ra.

Mỗi bước tính toán này lặp lại qua các thời điểm t trong chuỗi thời gian, với mô hình học từ dữ liệu trước đó để đưa ra dự đoán tại thời điểm hiện tại.

### Cách Thực Hiện Dự Đoán Số Lượng Khách Hàng Theo Ngày bằng RNN

Để dự đoán số lượng khách hàng vào ngày tiếp theo hoặc trong tương lai gần, cần thực hiện một số bước tiền xử lý dữ liệu và xây dựng mô hình RNN. Các bước thực hiện có thể được mô tả như sau:

#### Bước 1: Chuẩn Bị Dữ Liệu

Trước tiên, cần thu thập dữ liệu lịch sử về số lượng khách hàng đã đến thăm hoặc mua hàng vào mỗi ngày. Dữ liệu này có thể bao gồm:

* **Số lượng khách hàng** mỗi ngày (số liệu về lượng khách hàng thực tế).
* **Các yếu tố ngoại cảnh** như mùa vụ, ngày lễ, các chương trình khuyến mãi, các sự kiện đặc biệt có thể ảnh hưởng đến số lượng khách hàng.
* **Các thông tin liên quan đến marketing**: Chiến dịch quảng cáo, khuyến mãi, hay sự kiện đặc biệt trong ngày có thể làm thay đổi số lượng khách hàng.

#### Bước 2: Tiền Xử Lý Dữ Liệu

Dữ liệu cần được tiền xử lý để có thể đưa vào mô hình học máy. Các bước tiền xử lý bao gồm:

* **Xử lý missing values (giá trị thiếu):** Loại bỏ hoặc thay thế các giá trị thiếu trong dữ liệu. Điều này đảm bảo mô hình không bị thiếu sót thông tin quan trọng.
* **Tạo chuỗi thời gian:** Dữ liệu cần được chuyển đổi thành chuỗi thời gian, trong đó mỗi bước thời gian (mỗi ngày) sẽ chứa các giá trị liên quan đến số lượng khách hàng và các yếu tố ngoại cảnh.
* **Chia dữ liệu:** Dữ liệu thường được chia thành hai tập: tập huấn luyện (training set) và tập kiểm tra (test set). Một tỷ lệ phổ biến là 80/20, tức là 80% dữ liệu dùng để huấn luyện và 20% còn lại dùng để kiểm tra.

#### Bước 3: Chuẩn Hóa Dữ Liệu

Vì các giá trị trong dữ liệu có thể có đơn vị đo lường khác nhau (ví dụ: số lượng khách hàng có thể dao động lớn, trong khi các yếu tố khác như chương trình khuyến mãi có giá trị nhỏ), việc **chuẩn hóa dữ liệu** là rất cần thiết. Phương pháp chuẩn hóa phổ biến là **Min-Max Scaling**, giúp đưa tất cả các giá trị vào phạm vi từ 0 đến 1, giúp mô hình dễ dàng học và tối ưu hóa hơn.

#### Bước 4: Xây Dựng Mô Hình RNN

Khi dữ liệu đã được chuẩn bị, có thể xây dựng mô hình RNN để dự đoán số lượng khách hàng trong tương lai. Các bước xây dựng mô hình như sau:

1. **Khởi tạo mô hình RNN:** Mô hình có thể được xây dựng với các lớp ẩn (hidden layers) và các tham số như số lượng đơn vị (units), số lượng lớp (layers), và các tham số khác để cấu hình mô hình. Một mạng RNN cơ bản bao gồm các lớp đầu vào, lớp ẩn, và lớp đầu ra.
2. **Sử dụng LSTM (Long Short-Term Memory):** Nếu chuỗi dữ liệu có tính chất dài hạn và bạn muốn mô hình học các mối quan hệ lâu dài giữa các bước thời gian, bạn có thể thay thế lớp RNN cơ bản bằng lớp LSTM. Điều này giúp cải thiện khả năng dự đoán và xử lý các chuỗi thời gian dài hơn.
3. **Định dạng dữ liệu cho RNN:** Dữ liệu cần được định dạng theo kiểu **sequence-to-sequence** (seq-to-seq) hoặc **time series forecasting**, trong đó mỗi bước thời gian sẽ sử dụng một số lượng ngày trong quá khứ làm đầu vào để dự đoán số lượng khách hàng vào ngày tiếp theo. Đây là dạng dữ liệu phù hợp với mô hình học máy.
4. **Chọn hàm mất mát (loss function):** Hàm mất mát phổ biến cho các bài toán hồi quy là **Mean Squared Error (MSE)**, giúp tính toán sai số giữa giá trị dự đoán và giá trị thực tế. Ngoài ra, bạn có thể thử nghiệm với các hàm mất mát khác như **Mean Absolute Error (MAE)** để cải thiện kết quả.
5. **Tối ưu hóa mô hình:** Dùng các thuật toán tối ưu hóa như **Adam Optimizer** để huấn luyện mô hình, điều chỉnh các tham số và cải thiện độ chính xác của dự đoán.

#### Bước 5: Huấn Luyện Mô Hình

Sử dụng tập huấn luyện để huấn luyện mô hình. Trong quá trình huấn luyện, mô hình sẽ học được mối quan hệ giữa số lượng khách hàng trong quá khứ và các yếu tố ảnh hưởng đến sự thay đổi trong tương lai. Mô hình sẽ được huấn luyện qua nhiều **epoch** (vòng lặp), giúp tối ưu hóa các trọng số của mô hình để đạt được kết quả tốt nhất.

#### Bước 6: Đánh Giá Mô Hình

Sau khi huấn luyện, cần đánh giá mô hình bằng cách sử dụng tập kiểm tra. Các chỉ số đánh giá phổ biến bao gồm:

* **Mean Squared Error (MSE):** Đo lường sai số trung bình bình phương giữa giá trị dự đoán và giá trị thực tế. MSE càng nhỏ, mô hình càng chính xác.
* **Root Mean Squared Error (RMSE):** Căn bậc hai của MSE, giúp xác định mức độ sai lệch của mô hình. RMSE dễ dàng hiểu vì nó có cùng đơn vị đo lường với giá trị thực tế.
* **Mean Absolute Error (MAE):** Đo lường sai số tuyệt đối trung bình, thường được sử dụng khi không muốn quá chú trọng vào các sai số cực lớn.

#### Bước 7: Dự Đoán Số Lượng Khách Hàng

Khi mô hình đã được huấn luyện và đánh giá, có thể sử dụng mô hình để **dự đoán số lượng khách hàng** trong tương lai. Mô hình RNN sẽ sử dụng thông tin trong chuỗi thời gian (như số lượng khách hàng trong các ngày trước đó và các yếu tố ngoại cảnh) để đưa ra dự đoán cho ngày tiếp theo. Bạn có thể sử dụng mô hình để dự đoán số lượng khách hàng trong các ngày tiếp theo hoặc thậm chí trong nhiều tuần hoặc tháng.

### Lợi Ích Của Dự Đoán Số Lượng Khách Hàng Theo Ngày Bằng RNN

* **Tối ưu hóa kế hoạch nguồn lực:** Việc dự đoán chính xác số lượng khách hàng giúp doanh nghiệp lên kế hoạch tốt hơn về nhân sự, kho bãi, và hàng hóa.
* **Dự đoán xu hướng thay đổi:** RNN giúp nhận diện các xu hướng trong hành vi của khách hàng theo thời gian, ví dụ như những thời điểm nhu cầu tăng cao hoặc giảm mạnh.
* **Chạy thử các chiến lược marketing:** Mô hình dự đoán có thể giúp doanh nghiệp chạy thử các chiến lược marketing và dự đoán hiệu quả trong các điều kiện thị trường khác nhau.
* **Lập kế hoạch tài chính và bán hàng:** Doanh nghiệp có thể dự đoán số lượng khách hàng để điều chỉnh chính sách giá cả và tối ưu hóa các chiến lược bán hàng.

#### Lợi Ích Của Việc Ứng Dụng BI

* **Tối ưu hóa chiến lược marketing**: Dựa trên phân tích dữ liệu, doanh nghiệp có thể đưa ra các chiến lược marketing hiệu quả hơn, từ đó tăng trưởng doanh thu và giảm chi phí.
* **Cải thiện trải nghiệm khách hàng**: Nhờ vào việc phân tích hành vi và nhu cầu của khách hàng, doanh nghiệp có thể cá nhân hóa trải nghiệm khách hàng, từ đó tăng mức độ hài lòng và giữ chân khách hàng lâu dài.
* **Quản lý nguồn lực hiệu quả hơn**: Dự đoán nhu cầu khách hàng giúp doanh nghiệp phân bổ nguồn lực, nhân sự và vật tư hợp lý, giảm thiểu lãng phí và tối ưu hóa hoạt động kinh doanh.
* **Tăng trưởng bền vững**: Thông qua việc sử dụng BI, doanh nghiệp có thể đưa ra các quyết định dựa trên dữ liệu, giúp tăng trưởng bền vững và giảm thiểu các rủi ro.

# ỨNG DỤNG WEB VÀ KẾT QUẢ ĐẦU RA

## Dashboard

**Tổng quan doanh thu (Total Revenue):** Hiển thị tổng doanh thu thực hiện được ($233,650,194), giúp doanh nghiệp có cái nhìn toàn cảnh về hiệu suất tài chính.

**Phân tích doanh thu theo tháng (Month-Wise Revenue):** Biểu đồ xu hướng doanh thu theo từng tháng.

**Phân tích doanh thu theo độ tuổi (Age-Wise Sales Analysis):** Đánh giá mức tiêu thụ của khách hàng theo từng nhóm tuổi.

**Phân tích doanh thu theo vùng (Revenue per State & Region-Wise Revenue Share):** Thống kê doanh thu theo từng bang và khu vực.

**Phân tích doanh thu theo giới tính (Gender-Wise Sales Analysis):** So sánh mức độ mua sắm của nam và nữ theo từng danh mục sản phẩm.

**Tương quan chiết khấu và lượng đặt hàng (Quantity-Discount Correlation):** Phân tích mối quan hệ giữa tỷ lệ chiết khấu và số lượng sản phẩm bán ra.

Có thể xem các thông tin tương đương của từng loại sản phẩm bằng thang lựa chọn

### Month-Wise Revenue (Doanh thu theo tháng)

* **Mô tả:** Biểu đồ đường hiển thị doanh thu theo từng tháng, với đỉnh cao ở tháng 1 năm 2021 ($57.7M) và mức thấp ở tháng 4 năm 2021 ($4.4M).
* Ý nghĩa:
  + Doanh nghiệp có sự biến động lớn trong doanh thu theo từng tháng.
  + Đỉnh doanh thu vào tháng 1 có thể liên quan đến các chương trình khuyến mãi hoặc mùa cao điểm mua sắm.
  + Cần phân tích nguyên nhân giảm doanh thu vào các tháng thấp để cải thiện chiến lược kinh doanh trong tương lai.

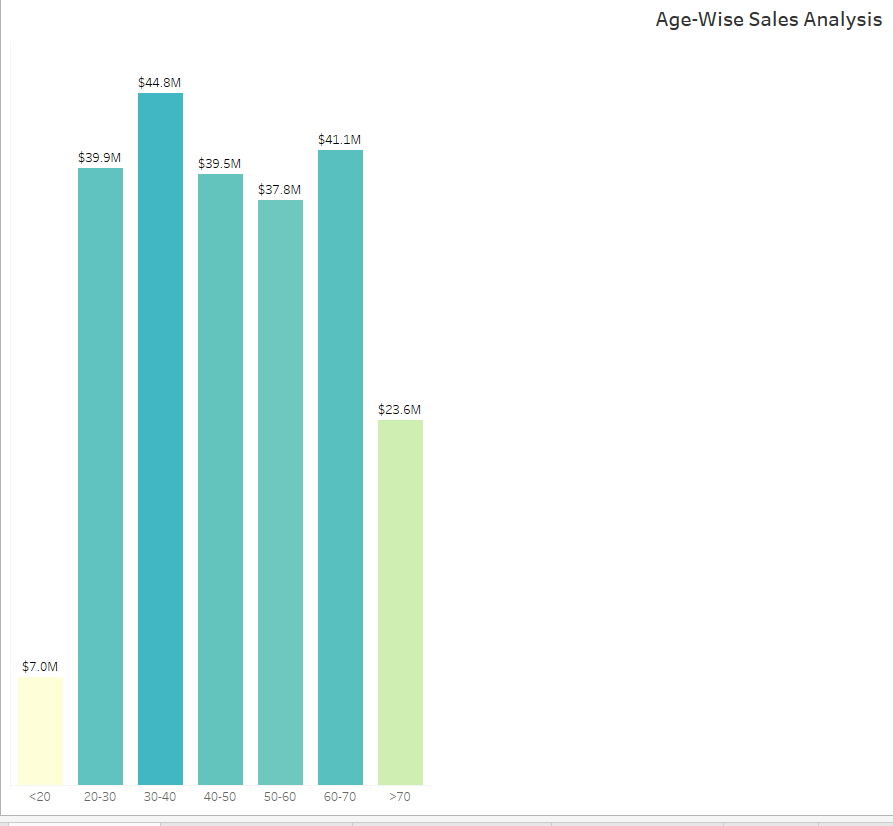
A graph of a graph with numbers and a line

Description automatically generated

Hình .: Doanh thu theo tháng

### Age-Wise Sales Analysis (Phân tích doanh thu theo độ tuổi)

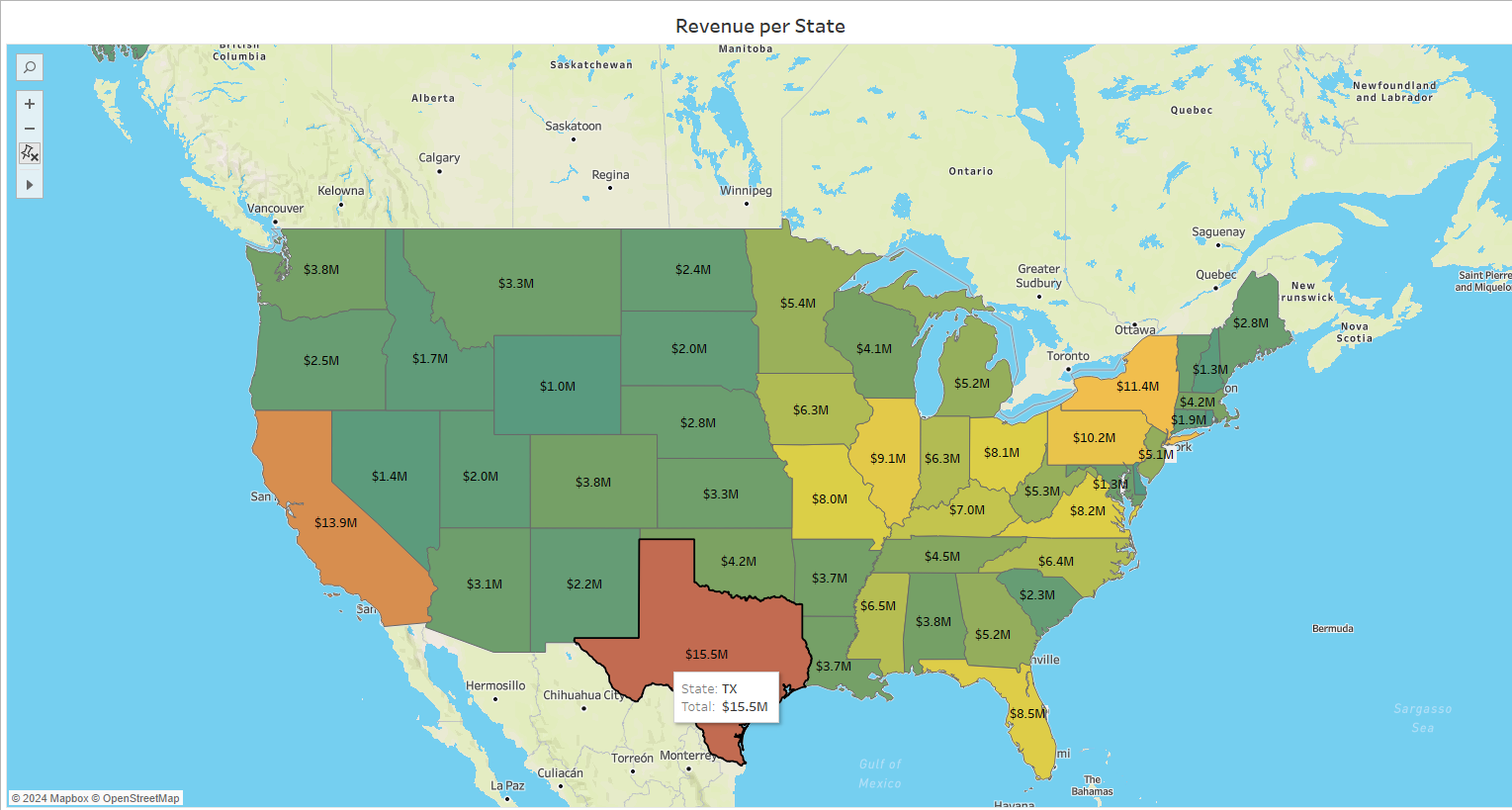
* **Mô tả:** Nhóm tuổi 30-40 có doanh thu cao nhất ($44.8M), tiếp theo là nhóm 60-70 ($41.1M). Nhóm dưới 20 tuổi có doanh thu thấp nhất ($7.0M).
* Ý nghĩa:
  + Nhóm tuổi 30-40 là đối tượng khách hàng trọng tâm, nên tập trung vào các sản phẩm và chiến lược marketing phù hợp với nhóm này.
  + Nhóm tuổi dưới 20 có doanh thu thấp, có thể do sản phẩm không phù hợp hoặc khả năng chi tiêu của nhóm này hạn chế.



Hình .: Phân tích bán hàng theo độ tuổi

### Revenu***e*** per State & Region-Wise Revenue Share (Doanh thu theo bang và khu vực)

* Mô tả:
  + Doanh thu cao nhất nằm ở bang Texas ($15.5M), trong khi một số bang khác có doanh thu thấp hơn $2M.
  + Phân chia doanh thu theo khu vực:
    - Miền Nam chiếm tỷ lệ lớn nhất (38.37%).
    - Miền Đông Bắc có tỷ lệ thấp nhất (17.10%).
* Ý nghĩa:
  + Texas là thị trường tiềm năng, cần duy trì và mở rộng chiến lược kinh doanh tại đây.
  + Doanh nghiệp có thể đầu tư thêm vào miền Đông Bắc để tăng tỷ lệ thị phần tại khu vực này.



Hình 7.3: Doanh thu trên mỗi tiểu bang

### Gender-Wise Sales Analysis (Phân tích doanh thu theo giới tính)

* Mô tả:
  + Nữ giới có doanh thu cao hơn ($66.27M) so với nam giới ($63.85M).
  + Các danh mục sản phẩm phổ biến:
    - Nữ: Thời trang nữ, điện thoại/máy tính bảng.
    - Nam: Thời trang nam, điện thoại/máy tính bảng.
* Ý nghĩa:
  + Cần tập trung phát triển các sản phẩm và chiến lược marketing phù hợp với từng giới tính.
  + Điện thoại/máy tính bảng là danh mục sản phẩm chủ lực cho cả hai giới, nên ưu tiên quảng bá.

A screenshot of a graph

Description automatically generated

Hình .: Phân tích bán hàng theo giới tính

### Quantity-Discount Correlation (Tương quan chiết khấu và lượng đặt hàng)

* **Mô tả:** Biểu đồ tán xạ thể hiện mối quan hệ giữa tỷ lệ chiết khấu và số lượng đơn hàng. Khi mức chiết khấu tăng, lượng đặt hàng có xu hướng tăng.
* Ý nghĩa:
  + Chiết khấu là yếu tố kích thích khách hàng mua sắm.
  + Doanh nghiệp nên tối ưu hoá các chương trình khuyến mãi để đạt hiệu quả doanh thu cao nhất.
  + Tuy nhiên, cần đảm bảo biên lợi nhuận không bị ảnh hưởng khi thực hiện chiết khấu.

A graph with blue dots

Description automatically generated

Hình .: Tương quan Số lượng - Giảm giá

### Kết luận

* **Tổng quan:** Dashboard này cung cấp các công cụ phân tích dữ liệu chi tiết, giúp doanh nghiệp hiểu rõ hơn về tình hình kinh doanh, từ đó đưa ra các quyết định chiến lược.
* Đề xuất:
  + Tăng cường tập trung vào các nhóm khách hàng tiềm năng (tuổi 30-40, khu vực miền Nam).
  + Phân tích nguyên nhân doanh thu thấp ở các tháng và khu vực cụ thể để cải thiện.
  + Tối ưu hóa chiến lược chiết khấu để tối đa hóa lượng đặt hàng mà không làm giảm biên lợi nhuận.

## Predicting Returning Customers

* **Mô tả:** Dự đoán tỷ lệ khách hàng sẽ quay lại hoặc rời đi trong một khoảng thời gian cụ thể (Quý 3 năm 2021 trong hình).
* Thông tin hiển thị:
  + **Tổng số khách hàng:** Tổng số khách hàng được phân tích (9,241 khách hàng).
  + **Tỷ lệ rời đi dự kiến:** Tỷ lệ khách hàng có khả năng rời đi (61.01%).
  + **Số lượng khách hàng có nguy cơ rời đi:** Số lượng cụ thể khách hàng có nguy cơ rời đi (5,638 khách hàng).
  + **Độ chính xác của mô hình:** Hiệu suất của mô hình dự đoán, được đánh giá bằng độ chính xác (83.77%).
* **Kết quả:**
  + Tổng số khách hàng trong dự đoán là 9,241.
  + Tỷ lệ rời đi dự kiến ở mức cao (61.01%), tương đương 5,638 khách hàng.
  + Độ chính xác của mô hình đạt 83.77%, cho thấy mô hình hoạt động hiệu quả và có độ tin cậy cao.
* **Ý nghĩa:**
  + Việc hơn 60% khách hàng có nguy cơ rời đi là một tín hiệu đáng báo động. Doanh nghiệp cần nhanh chóng triển khai các chiến lược giữ chân khách hàng.
  + Độ chính xác 83.77% đảm bảo các dự đoán là đáng tin cậy, giúp doanh nghiệp tập trung vào các khách hàng có nguy cơ cao.

A close up of a number

Description automatically generated

Hình .: Predicting Returning Customers

### Phân tích tỷ lệ quay lại theo các đặc điểm khách hàng

* **Theo giới tính:** So sánh tỷ lệ quay lại giữa nam và nữ.
  + **Kết quả:** Tỷ lệ quay lại giữa nam và nữ có sự chênh lệch nhẹ. Điều này chỉ ra rằng giới tính không phải là yếu tố quyết định lớn đến hành vi quay lại.
  + **Ý nghĩa:** Doanh nghiệp cần tập trung vào các yếu tố khác (như chi tiêu hoặc vùng địa lý) thay vì phân khúc khách hàng theo giới tính.
* **Theo độ tuổi:** Theo dõi khả năng quay lại theo từng nhóm tuổi.
  + **Kết quả:** Các nhóm tuổi có tỷ lệ quay lại khác nhau, với một số nhóm tuổi nhất định (ví dụ: 30-50) có vẻ trung thành hơn.
  + **Ý nghĩa:** Doanh nghiệp nên tập trung giữ chân khách hàng thuộc nhóm tuổi có tỷ lệ quay lại cao, đồng thời tìm cách cải thiện trải nghiệm cho các nhóm tuổi khác.
* **Theo vùng địa lý:** Đánh giá mức độ trung thành của khách hàng theo khu vực (Midwest, Northeast, South, West).
  + **Kết quả:** Tỷ lệ quay lại giữa các khu vực có sự khác biệt nhẹ, trong đó Midwest và West thường có tỷ lệ quay lại cao hơn.
  + **Ý nghĩa:** Doanh nghiệp có thể tập trung tăng cường chiến lược chăm sóc khách hàng tại các khu vực có tỷ lệ quay lại thấp (ví dụ: Northeast).

A graph with numbers and a red line

Description automatically generated with medium confidence

Hình .: Phân tích tỷ lệ quay lại theo các đặc điểm khách hàng

### Top các yếu tố ảnh hưởng đến hành vi rời đi

* **Mô tả:** Xác định các đặc trưng quan trọng nhất ảnh hưởng đến hành vi rời đi của khách hàng, bao gồm:
  + **Tổng chi tiêu (total\_spent):** Tổng số tiền khách hàng đã chi tiêu.
  + **Giá trị đơn hàng trung bình (avg\_order\_value):** Giá trị trung bình mỗi đơn hàng.
  + **Tỷ lệ chiết khấu trung bình (avg\_discount\_percent):** Tỷ lệ chiết khấu mà khách hàng đã nhận được.
* **Kết quả:**
  + Tổng chi tiêu (total\_spent) là yếu tố ảnh hưởng lớn nhất. Khách hàng chi tiêu ít thường có khả năng rời đi cao.
  + Giá trị đơn hàng trung bình (avg\_order\_value) cũng quan trọng. Những khách hàng có giá trị đơn hàng thấp hơn thường dễ rời đi.
  + Tỷ lệ chiết khấu trung bình (avg\_discount\_percent) cũng là một yếu tố tác động. Khách hàng nhận nhiều chiết khấu có thể hình thành kỳ vọng cao, dẫn đến khả năng rời đi khi không nhận được ưu đãi tương tự.
* **Ý nghĩa:**
  + Doanh nghiệp nên tập trung vào việc tăng giá trị đơn hàng trung bình và tổng chi tiêu của khách hàng thông qua chương trình khuyến mãi, bán chéo sản phẩm (cross-sell), hoặc nâng cấp sản phẩm/dịch vụ.
  + Tối ưu hóa các chương trình chiết khấu để vừa thu hút khách hàng nhưng không tạo kỳ vọng quá cao.

A white background with black lines

Description automatically generated

Hình .: Top các yếu tố ảnh hưởng đến hành vi rời đi

### Phân nhóm khách hàng có khả năng rời đi cao và thấp

* **Danh sách khách hàng có tỷ lệ rời đi trên 80%:** Hiển thị danh sách khách hàng có nguy cơ rời đi cao, kèm theo tên, email và xác suất rời đi.
  + **Kết quả:** Danh sách khách hàng cụ thể với tỷ lệ rời đi cao (>80%) được hiển thị chi tiết, kèm theo thông tin liên hệ.
  + **Ý nghĩa:** Doanh nghiệp cần ưu tiên thực hiện các chiến lược giữ chân nhóm khách hàng này, chẳng hạn:
    - Gửi email cá nhân hóa để khôi phục mối quan hệ.
    - Đưa ra chương trình khuyến mãi hoặc ưu đãi đặc biệt.

A screenshot of a computer

Description automatically generated

Hình .: Phân nhóm khách hàng có khả năng rời đi trên 80%

* **Danh sách khách hàng có tỷ lệ rời đi dưới 20%:** Hiển thị khách hàng có khả năng trung thành cao.
  + **Kết quả:** Nhóm khách hàng trung thành được xác định, với tỷ lệ rời đi rất thấp (<20%).
  + **Ý nghĩa:** Doanh nghiệp nên tiếp tục duy trì mối quan hệ tích cực với nhóm khách hàng này thông qua các chương trình khách hàng thân thiết hoặc dịch vụ chăm sóc đặc biệt.

A screenshot of a computer

Description automatically generated

Hình .: Phân nhóm khách hàng có khả năng rời đi dưới 20%

### Ý nghĩa của hệ thống đối với doanh nghiệp

* Tăng khả năng giữ chân khách hàng:
  + Dựa vào danh sách các khách hàng có nguy cơ rời đi cao, doanh nghiệp có thể triển khai các chiến lược cá nhân hóa để giảm tỷ lệ rời đi.
* Tối ưu hóa chiến lược kinh doanh:
  + Phân tích các yếu tố quan trọng như tổng chi tiêu và giá trị đơn hàng giúp doanh nghiệp điều chỉnh chiến lược bán hàng, khuyến mãi để tăng giá trị khách hàng.
* Hiểu rõ hành vi khách hàng:
  + Phân tích theo độ tuổi, giới tính và vùng địa lý giúp doanh nghiệp nhắm mục tiêu chính xác hơn trong các chiến dịch marketing.
* Tăng doanh thu dài hạn:
  + Việc giữ chân khách hàng trung thành (thay vì thu hút khách hàng mới) giúp doanh nghiệp tiết kiệm chi phí và tối ưu hóa lợi nhuận.

## Dự đoán số lượng khách hàng

Trang web "BI Dự đoán số lượng khách hàng" sử dụng mô hình học sâu (RNN - Recurrent Neural Network) để dự đoán số lượng khách hàng hàng ngày của từng cửa hàng. Công cụ này hỗ trợ doanh nghiệp phân tích và dự đoán xu hướng khách hàng, từ đó đưa ra các chiến lược kinh doanh phù hợp. Dưới đây là báo cáo chi tiết về chức năng, kết quả đầu ra của mô hình, và ý nghĩa của nó với doanh nghiệp.

### Các chức năng chính của trang web

* Mô tả:
  + Sử dụng mô hình RNN để dự đoán số lượng khách hàng hàng ngày của các cửa hàng dựa trên dữ liệu lịch sử.
  + Kết quả được hiển thị dưới dạng biểu đồ với ba đường:
    - **Actual Customers (Xanh dương):** Số lượng khách hàng thực tế.
    - **Predicted Customers (Đỏ):** Kết quả dự đoán từ mô hình RNN.
    - **Forecasted Customers (Xanh lá):** Dự báo số lượng khách hàng cho các ngày trong tương lai.
* Mục tiêu:
  + So sánh giá trị thực tế và giá trị dự đoán để đánh giá hiệu suất của mô hình.
  + Dự đoán xu hướng số lượng khách hàng trong tương lai để doanh nghiệp chuẩn bị nguồn lực.

A graph with red and blue lines

Description automatically generated

Hình .: Dự đoán số lượng khách hàng của ngày tiếp theo

### Kết quả đầu ra của mô hình RNN

#### Hiệu suất dự đoán số lượng khách hàng

* **Quan sát:**
  + Kết quả của mô hình RNN (đường màu đỏ) khá sát với dữ liệu thực tế (đường màu xanh dương) trong phần lớn các ngày.
  + Có một số thời điểm (như ngày cao điểm giữa tháng 9), mô hình thể hiện sự sai lệch giữa số lượng khách hàng thực tế và dự đoán, đặc biệt là khi có sự biến động đột ngột.
  + Kết quả dự báo cho tương lai (đường màu xanh lá) cho thấy xu hướng tăng nhẹ trong những ngày sắp tới.
* **Bình luận:**
  + Mô hình RNN hoạt động tốt trong việc dự đoán các xu hướng tổng quát và số lượng khách hàng trung bình, đặc biệt trong những ngày có biến động nhỏ.
  + Tuy nhiên, mô hình gặp khó khăn trong việc dự đoán các sự kiện bất thường (như ngày có lượng khách tăng đột biến), điều này có thể do thiếu thông tin về các yếu tố bên ngoài (chẳng hạn như chương trình giảm giá, thời tiết, hay sự kiện đặc biệt).

#### Độ chính xác của dự báo cho tương lai

* **Quan sát:**
  + Đường màu xanh lá (Forecasted Customers) cho thấy xu hướng tăng dần trong số lượng khách hàng trong những ngày tới.
  + Các giá trị dự báo khá ổn định, không có biến động bất thường, phù hợp với xu hướng tổng thể của dữ liệu quá khứ.
* **Bình luận:**
  + Dự báo tương lai của mô hình RNN có tính tin cậy cao, đặc biệt trong trường hợp xu hướng khách hàng không có sự thay đổi quá lớn.
  + Tuy nhiên, nếu có các yếu tố ảnh hưởng bên ngoài (như chiến dịch marketing hoặc ngày lễ), thì dự báo có thể cần được điều chỉnh để phản ánh thực tế.

### Ý nghĩa đối với doanh nghiệp

#### Tối ưu hóa nguồn lực

* Dự đoán chính xác số lượng khách hàng giúp doanh nghiệp chuẩn bị tốt hơn về nguồn lực như:
  + Nhân sự (điều chỉnh ca làm việc để tránh thiếu hụt hoặc dư thừa lao động).
  + Hàng tồn kho (đảm bảo lượng hàng đủ đáp ứng nhu cầu).
  + Dịch vụ (điều chỉnh không gian hoặc quy trình phục vụ để cải thiện trải nghiệm khách hàng).

#### Hỗ trợ lập kế hoạch kinh doanh

* Dự báo xu hướng khách hàng trong tương lai giúp doanh nghiệp:
  + Lên kế hoạch triển khai các chương trình khuyến mãi hoặc sự kiện vào những ngày thấp điểm để tăng doanh thu.
  + Tận dụng ngày cao điểm để tối đa hóa lợi nhuận, đồng thời giảm thiểu rủi ro quá tải.

#### Phát hiện và xử lý các bất thường

* Phân tích sự sai lệch giữa số lượng khách hàng thực tế và dự đoán giúp doanh nghiệp:
  + Nhận diện các yếu tố bất thường ảnh hưởng đến hành vi khách hàng (như sự kiện cạnh tranh, thời tiết, hay các yếu tố kinh tế).
  + Điều chỉnh mô hình dự đoán và chiến lược kinh doanh để cải thiện hiệu suất.

#### Tăng cường khả năng ra quyết định

* Xác định các xu hướng khách hàng trong dài hạn giúp doanh nghiệp:
  + Đưa ra các quyết định chiến lược dựa trên dữ liệu (data-driven decisions).
  + Tăng khả năng cạnh tranh trên thị trường bằng cách đáp ứng nhu cầu khách hàng hiệu quả hơn.

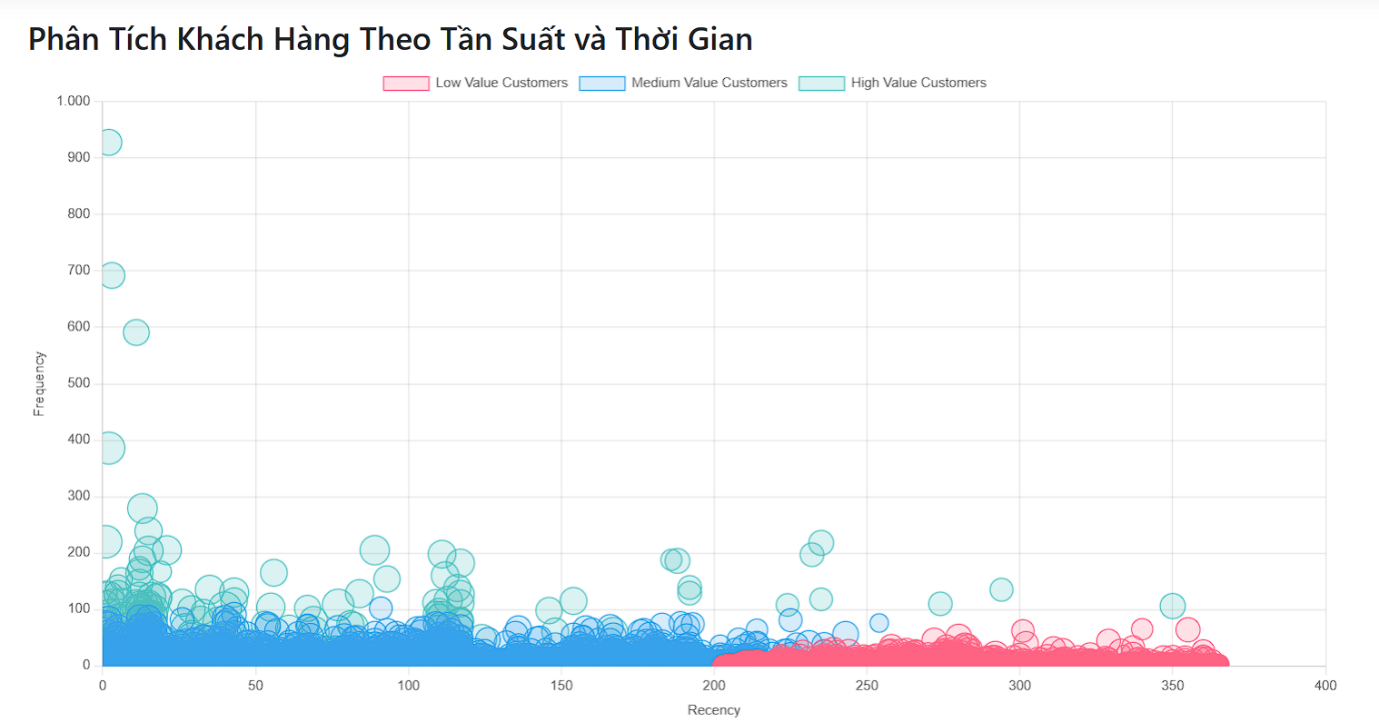
## Phân khúc khách hàng

Trang web "BI Phân khúc khách hàng" sử dụng mô hình K-Means để phân tích và phân khúc khách hàng dựa trên dữ liệu hành vi mua sắm (Recency, Frequency, Monetary - RFM). Qua đó, doanh nghiệp có thể hiểu rõ hơn về các nhóm khách hàng của mình và tối ưu hóa chiến lược kinh doanh phù hợp.

### Các chức năng chính của trang web

#### Phân khúc khách hàng theo mô hình RFM

* **Mô tả:**
  + Sử dụng ba yếu tố RFM để phân tích khách hàng:
    - **Recency (R):** Thời gian gần nhất khách hàng thực hiện giao dịch.
    - **Frequency (F):** Số lần khách hàng thực hiện giao dịch trong một khoảng thời gian nhất định.
    - **Monetary (M):** Tổng giá trị giao dịch của khách hàng.
  + Dựa trên mô hình K-Means, khách hàng được chia thành ba nhóm:
    - Low Value Customers (Nhóm giá trị thấp - màu đỏ).
    - Medium Value Customers (Nhóm giá trị trung bình - màu xanh dương).
    - High Value Customers (Nhóm giá trị cao - màu xanh lục nhạt).
* **Hiển thị:**
  + Biểu đồ phân tán thể hiện mối quan hệ giữa Recency và Frequency, phân nhóm khách hàng bằng các màu sắc khác nhau.
  + Danh sách chi tiết gồm tên khách hàng, giá trị RFM, và nhóm phân khúc.



Hình .: Phân tích khách hàng theo tần suất và thời gian

#### Tỷ lệ các nhóm khách hàng

* Mô tả:
  + Biểu đồ tròn hiển thị tỷ lệ phần trăm của từng nhóm phân khúc khách hàng (Low, Medium, High Value).
  + Cung cấp cái nhìn tổng quan về sự phân bố khách hàng trong từng nhóm.

A pie chart with a black box and a black tag

Description automatically generated with medium confidence

Hình .: Tỷ lệ các nhóm khách hàng

#### So sánh giá trị trung bình của các nhóm khách hàng

* Mô tả:
  + Biểu đồ cột thể hiện sự khác biệt về giá trị trung bình của các yếu tố Recency, Frequency, và Monetary giữa các nhóm khách hàng.
  + Phân tích sâu hơn giúp nhận diện đặc điểm của từng nhóm.

A white and blue graph

Description automatically generated

Hình .: So sánh các giá trị trung bình các nhóm khách hàng

A screenshot of a graph

Description automatically generated

Hình .: So sánh các giá trị trung bình các nhóm khách hàng

A graph with red squares

Description automatically generated

Hình .: So sánh các giá trị trung bình các nhóm khách hàng

#### Danh sách chi tiết khách hàng

* Mô tả:
  + Hiển thị thông tin chi tiết của từng khách hàng, bao gồm:
    - Tên khách hàng.
    - Giá trị Recency, Frequency, và Monetary.
    - Nhóm phân khúc của khách hàng.
  + Chức năng lọc theo nhóm phân khúc để dễ dàng truy xuất dữ liệu cụ thể.

A screenshot of a computer

Description automatically generated

Hình .: Phân loại và phân khúc khách hàng

### Kết quả đầu ra của mô hình K-Means

#### Hiệu quả phân khúc

* **Quan sát:**
  + Biểu đồ phân tán cho thấy các nhóm khách hàng được phân tách rõ ràng theo các giá trị RFM:
    - Nhóm **High Value Customers** tập trung ở phía dưới bên phải (tần suất mua hàng cao, giá trị giao dịch lớn, và thời gian giao dịch gần nhất).
    - Nhóm **Low Value Customers** nằm rải rác ở phía trên bên trái (tần suất mua thấp, giá trị giao dịch nhỏ, và thời gian giao dịch xa).
    - Nhóm **Medium Value Customers** nằm giữa hai nhóm còn lại.
* **Bình luận:**
  + Mô hình K-Means đã phân nhóm khách hàng khá hiệu quả, giúp doanh nghiệp dễ dàng nhận diện các nhóm khách hàng quan trọng.
  + Tuy nhiên, để tăng độ chính xác, mô hình có thể được tinh chỉnh thêm bằng cách thử nghiệm với số cụm K khác nhau và kết hợp thêm các thông tin bổ sung (như hành vi truy cập website, độ hài lòng khách hàng...).

#### Tỷ lệ phân bố các nhóm khách hàng

* **Quan sát:**
  + Biểu đồ tròn cho thấy tỷ lệ nhóm khách hàng giá trị thấp (Low Value Customers) chiếm tỷ lệ cao nhất.
  + Nhóm khách hàng giá trị trung bình và cao chiếm tỷ lệ nhỏ hơn.
* **Bình luận:**
  + Sự chênh lệch tỷ lệ giữa các nhóm cho thấy doanh nghiệp có một lượng lớn khách hàng ít mua sắm hoặc giá trị giao dịch thấp.
  + Doanh nghiệp cần tập trung vào việc chuyển đổi nhóm khách hàng Low Value sang Medium hoặc High Value thông qua các chiến lược giữ chân và tăng giá trị giao dịch.

#### Khác biệt về giá trị trung bình giữa các nhóm

* **Quan sát:**
  + Biểu đồ cột cho thấy sự chênh lệch đáng kể về giá trị Monetary giữa nhóm khách hàng High Value và hai nhóm còn lại.
  + Nhóm High Value có giá trị Frequency cao hơn, cho thấy họ là những khách hàng thường xuyên và mang lại giá trị lớn.
* **Bình luận:**
  + Nhóm High Value không chỉ đóng góp nhiều về doanh thu mà còn thể hiện sự trung thành cao. Doanh nghiệp cần xây dựng các chương trình chăm sóc đặc biệt cho nhóm này để duy trì mối quan hệ lâu dài.
  + Nhóm Low Value cần được kích thích để tăng Frequency và Monetary thông qua các chiến lược khuyến mãi, giảm giá, hoặc ưu đãi dành riêng.

### Ý nghĩa của phân khúc khách hàng đối với doanh nghiệp

#### Tối ưu hóa chiến lược marketing

* Doanh nghiệp có thể xây dựng các chiến lược marketing riêng biệt cho từng nhóm khách hàng:
  + High Value Customers:
    - Tăng cường chương trình khách hàng thân thiết, ưu đãi đặc biệt, và trải nghiệm cá nhân hóa để duy trì sự trung thành.
  + Medium Value Customers:
    - Đẩy mạnh các chiến dịch upsell hoặc cross-sell để tăng giá trị giao dịch.
  + Low Value Customers:
    - Thu hút sự quan tâm của nhóm này bằng các chương trình khuyến mãi, quà tặng, hoặc chiến lược tái kích hoạt.

#### Tăng doanh thu và lợi nhuận

* Việc tập trung vào nhóm High Value giúp doanh nghiệp tối đa hóa lợi nhuận từ nhóm khách hàng hiện tại.
* Đồng thời, việc cải thiện nhóm Low Value sẽ giúp mở rộng tập khách hàng giá trị cao trong tương lai.

#### Cải thiện trải nghiệm khách hàng

* Phân khúc khách hàng giúp doanh nghiệp hiểu rõ nhu cầu và hành vi của từng nhóm, từ đó cải thiện chất lượng dịch vụ và trải nghiệm cá nhân hóa.
* Điều này không chỉ tăng sự hài lòng mà còn giúp tăng khả năng giữ chân khách hàng.

#### Định hướng chiến lược dài hạn

* Phân tích tỷ lệ và đặc điểm của từng nhóm khách hàng giúp doanh nghiệp định hướng chiến lược sản phẩm, giá cả, và phân phối phù hợp với từng phân khúc.
* Doanh nghiệp có thể ưu tiên đầu tư vào các nhóm khách hàng mang lại giá trị cao nhất, đồng thời nâng cao hiệu quả hoạt động kinh doanh.

### Kết luận

Trang web "BI Phân khúc khách hàng" là một công cụ mạnh mẽ giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình thông qua mô hình K-Means. Kết quả phân khúc không chỉ hỗ trợ doanh nghiệp tối ưu hóa chiến lược marketing mà còn tăng cường khả năng ra quyết định dựa trên dữ liệu.

Doanh nghiệp cần tiếp tục cải thiện mô hình phân tích bằng cách thu thập thêm dữ liệu và tích hợp yếu tố bên ngoài để nâng cao độ chính xác và hiệu quả.

# TÀI LIỆU THAM KHẢO

1. Kaggle. (2020). Sales Data FY 2020-2021. Retrieved from <https://www.kaggle.com>

2. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

5. Tableau Software. (2024). Visual Analytics Best Practices. Tableau Public. Retrieved from <https://www.tableau.com>

6. Flask. (2024). Flask Documentation. Retrieved from https://flask.palletsprojects.com

7. MongoDB Inc. (2024). MongoDB Documentation. Retrieved from <https://www.mongodb.com/docs>

8. Suresh, K. (2021). Analyzing Customer Behavior using RFM Analysis. International Journal of Data Science and Analytics, 14(3), 145–156.